Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Andrej Čopar

Modeliranje 3D struktur interakcij med proteini in RNA

magistrsko delo

študijski program druge stopnje
računalništvo in informatika

Mentor: doc. dr. Tomaž Curk

Ljubljana, 2014

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Izjava o avtorstvu magistrskega dela

Spodaj podpisani Andrej Čopar, z vpisno številko **63090054**, sem avtor magistrskega dela z naslovom:

Modeliranje 3D struktur interakcij med proteini in RNA

S svojim podpisom zagotavljam, da:

- sem magistrsko delo izdelal samostojno pod mentorstvom doc. dr. Tomaža Curka,
- so elektronska oblika magistrskega dela, naslov (slov., angl.), povzetek (slov., angl.) in ključne besede (slov., angl.) identični s tiskano obliko magistrskega dela,
- soglašam z javno objavo elektronske oblike magistrskega dela v zbirki "Dela FRI".

V Ljubljani, 1. septembra 2014 Podpis avtorja:

Zahvala

Zahvaljujem se mentorju, doc. dr. Tomažu Curku, za vodenje in nepogrešljivo strokovno usmeritev pri izdelavi magistrskega dela. Iskreno se zahvaljujem tudi družini za vso podporo.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
AUC	area under roc curve	površina pod krivuljo ROC
CA	classification accuracy	klasifikacijska točnost
MCC	Matthew's correlation coefficient	korelacijski koeficient Matthew
NBC	naive Bayes classifier	naivni Bayesov klasifikator
NMR	nuclear magnetic resonance spectroscopy	jedrska magnetna resonančna spektroskopija
PSSM	position specific scoring matrices	položajne ocenjevalne matrike
RBD	RNA binding domain	RNA-vezavna domena
RBP	RNA binding proteins	RNA-vezujoči proteini
RF	random forest	metoda naključnih gozdov
RNA	ribonucleic acid	ribonukleinska kislina
SVM	support vector machine	metoda podpornih vektorjev
RMSD	root mean square distance	povprečna kvadratna razdalja
Å	angstrom (0.1 nm)	angstrom (0.1 nm)

Povzetek

Interakcije med proteini in RNA imajo ključno vlogo pri velikem številu celičnih procesov. Eksperimentalna analiza 3D struktur molekul je počasna in zahtevna, zato obstaja velika potreba po računskih metodah, ki uspešno napovedujejo mesta ter strukturo molekul v interakciji. V magistrskem delu smo definirali vrsto značilk, ki opisujejo lokalne lastnosti interakcij protein-RNA, na podlagi podatkov o 3D strukturah molekul protein-RNA. Razvili smo metodo, ki združuje strojno učenje in optimizacijski postopek za napovedovanje mesta interakcij med proteinom in RNA. Napovedi strojnega učenja se uporabijo za določanje začetnega stanja optimizacije. Optimizacijski postopek nato uporabi ocenjevalne funkcije osnovane na porazdelitvi 3D strukturnih značilk in tako predlaga najverjetnejšo pozicijo molekule RNA. Predlagani napovedni model dosega natančnost, ki je primerljiva z uspešnostjo najboljših obstoječih metod.

Ključne besede:

bioinformatika, interakcije protein-RNA, strukturna analiza, napovedni model, kombinatorična optimizacija, umestitev molekul

Abstract

Protein-RNA interactions have an essential role in many cellular processes. Experimental analysis of 3D molecular structure is slow and difficult process. Consequently, computational methods, which successfully predict interaction sites and molecular conformations are needed. In this thesis we have defined a number of attributes to describe local properties of protein-RNA interactions using data on 3D structure of protein-RNA molecules. We have implemented a method that uses machine learning and optimization algorithm for prediction of protein-RNA interaction sites. Machine learning predictions are used to generate initial positions for optimization. Optimization algorithm uses scoring functions based on the distribution of 3D structural attributes to identify most likely positions of the RNA molecule interacting with a given protein. The accuracy of the proposed prediction model is comparable to results obtained with best existing methods.

Keywords:

bioinformatics, protein-RNA interactions, structural analysis, prediction model, combinatorial optimization, molecular docking

Poglavje 1

Uvod

Interakcije protein-RNA bistveno vplivajo na različne celične procese. Razumevanje molekularnih mehanizmov, ki vplivajo na interakcije protein-RNA, je ključnega pomena v biologiji, pomembno pa je tudi v industrijskih ter medicinskih aplikacijah.

V zadnjih letih je število podatkov o 3D strukturah proteinov in RNA bliskovito naraslo, a le malo metod te podatke izkorišča. Obstoječe raziskave na tem področju po večini temeljijo na analizi zaporedja in analizi 3D podatkov. Še vedno primanjkuje splošnih javno dostopnih orodij, zato smo si zadali naslednje cilje:

1. Določitev in ovrednotenje značilk za opisovanje prostorskih relacij med aminokislinami in nukleotidi v interakciji.

2. Gradnja modela za napovedovanje mest interakcije.

3. Implementacija umestitvenega algoritma za napovedovanje lokalne 3D strukture RNA.

V tem delu smo analizirali preko 200 struktur protein-RNA in zgradili model za napovedovanje mest interakcij protein-RNA. Modela smo vrednotili s trikratnim prečnim preverjanjem na podatkih o mestih interakcije na proteinu in podatkih o mestih interakcije na RNA. Predlagali smo optimizacijski postopek umestitve protein-RNA, ki uporablja statistične podatke

3D značilnosti interakcij in ocenjevalno funkcijo za vrednotenje optimalnih strukturnih kompleksov. Predlagane rešitve smo vrednotili z razdaljo do pravilne lege mesta RNA v interakciji.

V drugem poglavju predstavimo proteine, ki vežejo RNA in opišemo znane metode, ki za napovedovanje interakcij protein-RNA uporabljajo podatke o zaporedju ter metode, ki za napovedovanje uporabljajo tudi strukturne podatke.

V tretjem poglavju opišemo metodo umestitve protein-RNA (angl. docking) in možne funkcije za ocenjevanje napovedne uspešnosti ter obstoječe metode na tem področju.

V četrtem poglavju predlagamo model za napovedovanje mest interakcij protein-RNA na podlagi podatkov o 3D strukturah kompleksov protein-RNA. Za opis interakcij smo razvili vrsto značilk. Poleg podatka o tipu aminokisline in polarnosti smo razvili metode za opis lokalnih strukturnih značilnosti proteinov: koti in dolžine aminokislin, bližine sosednjih interakcij ter število okoliških aminokislin. Uporabimo tudi verjetnosti pojavitve parov aminokislina-nukleotid in orientacijo atomov znotraj aminokisline ter nukleotida. Zgradimo model strojnega učenja z različnimi klasifikatorji in predstavimo metrike za vrednotenje modela.

V petem poglavju predlagamo in opišemo metodo umestitve RNA na protein. Metoda uporablja optimizacijski iskalni algoritem in ocenjevalno funkcijo, ki je osnovana na uteženih verjetnostih, pridobljenih iz strukturnih podatkov. V optimizacijskem postopku generira najverjetnejše strukture RNA v interakciji s proteinom. Za dani protein določimo lokalno 3D strukturo RNA molekule in mesto na proteinu, ki vstopa v interakcijo.

V šestem poglavju predstavimo verjetnosti interakcij za posamezne aminokisline in nukleotide. Vrednotimo naš napovedni model in porazdelitve nekaterih značilk. Predstavimo tudi rešitve umestitvenega algoritma in njegovo točnost.

V zadnjem poglavju opišemo prispevke magistrske naloge, uspešnost izpolnjenih ciljev ter nadaljnje delo.

Poglavje 2

Interakcije protein-RNA

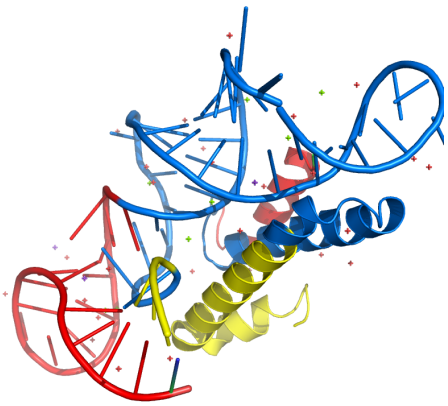
Proteini v interakciji z RNA so pri evkariontih kritične komponente genskih izraznih poti. Vključeni so v različne in pomembne celične procese prek njihove sposobnosti regulacije nastanka, stabilnosti, transporta in lokalizacije RNA. Razumevanje molekularnih mehanizmov interakcij protein-RNA in tvorbe kompleksnih struktur je velik izziv v strukturni biologiji. Poleg tega je to tudi ključni cilj za medicinske in farmacevtske namene, kot na primer odkrivanje zdravil, zlasti če upoštevamo, da so interakcije protein-RNA pogosto vključene v replikacijo in prepoznavo virusov [35].

Eksperimentalno določanje kompleksov protein-RNA z rentgensko kristalografijo in jedrsko magnetno resonančno spektroskopijo (NMR) je naporen in dolgotrajen proces. Interakcije protein-RNA lahko poskusimo napovedati z računskimi metodami. Čeprav je napoved manj natančna kot eksperimentalno opazovanje, so lahko računske napovedi dovolj natančne, da vodijo eksperimente in funkcijske hipoteze za identifikacijo posameznih aminokislin ali nukleotidov. V tem poglavju opišemo vlogo proteina in RNA v celicah ter opišemo tipična vprašanja, s katerimi se srečujemo pri problemu napovedi interakcij. V nadaljevanju poglavja bomo predstavili sorodna dela na področju interakcij protein-RNA. S tega področja sta glavna dva načina napovedovanje mest interakcij na podlagi zaporedij in na podlagi 3D struktur kompleksov protein-RNA.

2.1 RNA-vezavni proteini

Življenje tvori množica kompleksnih in med seboj povezanih interakcij, ki so odvisne od prisotnosti proteina, ne samo kot katalizatorja kemijskih reakcij, ampak tudi njegove vloge kot strukturna molekula in kot prenašalca molekul. Protein je sestavljen iz zaporedja aminokislin. Celica uporablja 20 različnih aminokislin za kodiranje proteinov, ki se med seboj razlikujejo po svojih lastnostih.

Proteini so pogosto v interakciji z ostalimi komponentami celice, majhnimi molekulami, nukleinskimi kislinami, membranami in drugimi proteini, ki skupaj tvorijo zapletene komplekse. Ena izmed teh molekul je tudi RNA (ribonukleinska kislina), ki je nujna za vrsto funkcij v celici. Skupna točka vrste funkcij, ki jo opravlja RNA, so interakcije s proteini. RNA skupaj s proteinom tvori del ribosoma in kaže katalizatorske sposobnosti ter opravlja spajanje molekul. Primer ribosomalnega fragmenta, ki je v interakciji z RNA, je prikazan na sliki 2.1.



Slika 2.1: Ribosomalni fragment z zapisom 1f7y v PDB.

Proteine, ki so v interakciji z RNA, imenujemo tudi RNA-vezavni proteini (RBP). Odgovorni so za transkripcijo, replikacijo, prenos, predelavo in uravnavo RNA v celicah. Napake v RNA-vezavnih proteinih so povezane s

številnimi boleznimi, ki segajo od nevroloških motenj do raka. RNA-vezavni proteini tipično vsebujejo različne strukturne motive, denimo RNA prepoznavni motiv (RRM), K-homologne domene (KH) in dvojno vijačne RNA- vezavne domene. Pri nekaterih proteinskih domenah skoraj vsi proteini kažejo RNA-vezavno aktivnost. Primer takih domen sta RRM in dsRBD domena. V drugih družinah je RNA-vezavna lastnost prisotna le deloma, na primer v encimskih družinah Rossman-fold metil-transferaze (RFM) [36]. Sorodne domene lahko vežejo RNA, DNA, proteine in druge substrate. Številčnost in raznolikost RNA-vezavnih proteinov je povezana s kompleksnostjo organizma.

2.2 Napovedovanje interakcij protein-RNA

Kljub pomembni funkcijski vlogi so interakcije protein-RNA še vedno slabše raziskano v primerjavi s precej bolj raziskanim področjem protein-DNA kompleksov. Za razliko od DNA, ki je ponavadi v obliki dvojne vijačnice, se RNA v interakcijah pojavlja tudi z eno vijačnico, lahko pa tudi s krajšimi odseki ali posameznimi nukleotidi. Na sliki 2.2 je prikazan primer interakcije protein-RNA na strukturi s PDB zapisom 1d6k.

Računsko napovedovanje vezav RNA opredeli tri povezane probleme:

1. ali je dan protein v interakciji z RNA,

2. če je, katere aminokisline v verigi proteina so neposredno vključene v interakcijo z RNA,

3. kakšna je struktura kompleksa protein-RNA.

Fosfatna osnovna veriga RNA je negativno nabita in je pogosto v interakciji s pozitivno nabitimi aminokislinami, na primer arginin in lizin. Vendar pa niso vse pozitivno nabite aminokisline vključene v vezavo z RNA. Pravzaprav večina proteinov vključuje površinsko izpostavljene in pozitivno nabite aminokisline, vse pa zagotovo niso v interakciji z RNA. Lahko so vključene

Slika 2.2: Primer interakcije protein-RNA z zapisom 1d6k v PDB.

v interakcijo z drugimi anionskimi ligandi, predvsem z DNA, ki ima zelo podobno osnovno verigo, lahko pa tudi za tvorjenje solnih mostov, katalize in v drugih funkcijah. Relativno razmerje med pozitivno in negativno nabitimi aminokislinami, ki ga lahko izračunamo iz zaporedja proteina, je slab prediktor vezave RNA. Obstajajo negativno nabiti proteini, ki vežejo anionske ligande vključno z nukleinskimi kislinami. Računske metode so bile razvite za prepoznavo RNA-vezavnih proteinov, osnovanih predvsem na naboju. Nekateri izmed njih uporabijo druge lastnosti zaporedja, na primer sestava aminokislin, van der Waals jakost in polarnost [36].

Za napovedovanje interakcij protein-RNA so Cheng et. al. uporabili položajno specifične ocenjevalne matrike (PSSM) [5], medtem ko so Kumar et. al., Yu et. al. ter Fujishima et. al. uporabili aminokislinsko sestavo in frekvence kot vektorje značilk za uporabo v SVM [57, 11]. Napovedni model razlikuje DNA ali RNA- vezavne proteine od proteinov, ki imajo druge funkcije. Za proteine z znanimi strukturami je Mandel-Gutfreund et. al. razvil metodo za identifikacijo območij pozitivno nabitih aminokislin na površini in razlikovanje med različnimi tipi proteinov, ki vežejo DNA in RNA [41].

Eksperimentalno določanje kompleksov protein-RNA je žal počasen in težaven proces. Zaradi pomankanja eksperimentalno določenih struktur kom-

pleksov protein-RNA so računske metode za napovedovanje kompleksov protein-RNA pomembne za preučevanje interakcij protein-RNA. Obstaja mnogo nizkoresolucijskih eksperimentalnih podatkov, ki določijo komponente v interakciji in jih pogosto povežejo z določenim funkcijskim stanjem. Ti podatki so lahko uporabljeni na področju bioinformatike za napovedovanje struktur. Kljub temu, da je metodologija za napovedovanje in modeliranje proteinov in protein-protein kompleksov dobro uveljavljena, obstaja le malo metod za napovedovanje in modeliranje RNA struktur in interakcij protein-RNA. V nadaljevanju poglavja bomo opisali obstoječe metode bioinformatike za napovedovanje struktur kompleksov protein-RNA.

2.3 Napovedovanje mest interakcije na podlagi zaporedja proteina

Napovedovanje zaporedij aminokislin, ki so v interakciji z RNA, se zanaša na uporabo metod strojnega učenja recimo nevronske mreže, skritih markovskih modelov (HMM) in metod podpornih vektorjev (SVM). Algoritmi tipično upoštevajo fiziokemične lastnosti aminokislin, zlasti naboj, hidrofobnost in napovedane lastnosti kot so dostopnost topila in sekundarne strukture, pogosto tudi ohranjanje zaporedja v lokalnem kontekstu. V nadaljevanju bomo opisali nekaj javno dostopnih orodij za napovedovanje RNA-vezavnih mest, ki uporabljajo zaporedje aminokislin brez dodatnih strukturnih podatkov za napovedovanje mest interakcij protein-RNA. Algoritmi SCRPRED, PPRINT in PRINTR uporabljajo evolucijske informacije o zaprodju proteina, medtem ko RPISeq uporablja zreducirano abecedo za opis zaporedij proteina in RNA. Za prepoznavo RBP in parov protein-RNA (RPISeq) so uporabljene metode strojnega učenja, kot so SVM, nevronske mreže (NN) in naključni gozdovi (RF).

RPISeq metoda sestoji iz dveh klasifikatorjev: RPISeq-SVM in RPISeq-RF [32]. V obeh algoritmih je uporabljenih 343 lastnosti za kodiranje pro-

tein zaporedja in 256 lastnosti za kodiranje zaporedja RNA. Proteini so kodirani z uporabo predstavitve trojic, pri katerih je vsaka izmed 20 aminokislin razvrščena v eno izmed 7 skupin. Torej vsako aminokislinso predstavlja 343 dimenzionalni vektor, pri kateri je vsak element vektorja normalizirana frekvenca ustrezne trojke v zaporedju. Zaporedja RNA so kodirana s 4- terkami v 256 dimenzionalni vektor, kjer vsaka lastnost predstavlja normalizirano frekvenco, s katero se dana 4-terka pojavlja v RNA zaporedju. Avtorji so metodo testirali na množicah iz PRIDB baze interakcij protein- RNA in dosegli klasifikacijsko točnost 76 %–90 % z metodo naključnih gozdov in 73 %–87 % z metodo SVM. Metoda doseže površino pod krivuljo ROC 0.92-0.97 (RF) in 0.81-0.85 (SVM).

SRCPRED uporablja nevronske mreže za napovedovanje aminokislin v interakciji z RNA z uporabo zaporedja (sestava aminokislin, GAC ocena) in evolucijsko informacijo (lokacijsko specifično ocenjevalno matriko, PSSM ocenjevanje) [10]. Za učenje nevronskih mrež so avtorji uporabili matriko zaporedne soseščine in matrike parov protein-RNA. Površina pod krivuljo ROC sega od 0.61 do 0.84 za štiri testirane funkcijske razrede RNA (virusna RNA, mRNA, tRNA in rRNA).

PPRINT kombinira evolucijsko informacijo in SVM za napovedovanje mest RNA-vezavnih mest v protein sekvencah [20]. Prvotna podatkovna množica vključuje 86 proteinov, ki so v interakciji z RNA, pridobljenih iz kompleksov protein-RNA. Evolucijska informacija je dobljena iz PSSM ocenjevanja, ki ga generira PSI-BLAST med iskanjem neredundančne baze proteinskih zaporedij. PPRINT doseže klasifikacijsko točnost 76 %, ko napoveduje aminokisline, ki so v kontaktu z RNA in MCC (korelacijski koeficient Matthew) 0.45.

PRINTR uporablja SVM in položajno specifične ocenjevalne matrike za napovedovanje interakcij protein-RNA [53]. Metoda uporablja večkratno poravnavo zaporedja, dostopnost topila in sekundarno strukturo. Me-

toda dosega površino pod krivuljo ROC velikosti 0.83 pri napovedi na nehomolognem naboru kompleksov (homolognost zaporedja manj kot 30 %).

BindN uporablja SVM za napovedovanje RNA vezavnih aminokislin na osnovi pKa vrednosti, hidrofobnosti in molekularne mase aminokislin. Zmožen je tudi napovedovanja DNA vezavnih aminokislin [54].

PRBR kombinira algoritem obogatenih naključnih gozdov (ERF) s hibridnim vektorjem značilk, sestavljenim iz napovedane sekundarne strukture, konzervacijske informacije fiziokemičnih lastnosti aminokislin in informacije o odvisnosti aminokislin glede na polarni naboj ter hidrofobnosti zaporedja v proteinu [25].

PiRaNhA uporablja SVM klasifikator za napovedovanje aminokislin v interakciji z RNA ali DNA. Klasifikator uporablja položajno specifične ocenjevalne matrike, verjetnostne porazdelitve aminokislinskih interakcij, dostopno površino in hidrofobnost.

2.4 Napovedovanje mest interakcije na podlagi strukture proteina

Dostopnost 3D strukture proteina lahko močno izboljša kvaliteto napovednega modela mest interakcij. Mesta interakcij so običajno sestavljena iz površinsko izpostavljenih aminokislin, ki so si blizu v prostoru, ampak ne nujno tudi v zaporedju. RNA-vezavna mesta lahko pogosto prepoznamo kot pozitivno nabita površinska mesta, katerih oblika je združljiva z obliko negativno nabite verige RNA. Poleg tega vizualni pregled omogoča iskanje delov z aromatičnimi in hidrofobnimi odseki, ki so lahko vključeni v zlaganje interakcij z bazami enostranske RNA. Napovedne metode, osnovane na prostorski strukturi, lahko izkoristijo isto informacijo kot metode zaporedja, le da zamenjajo napovedane lokalne strukturne lastnosti z opaženimi

lastnostmi iz 3D strukture. Poleg tega lahko uporabijo bolj globalne lastnosti, ki so dostopne le na 3D nivoju, kot na primer oblika površine, porazdelitev elektrostatičnega potenciala in prostorska bližina aminokislin z določenimi lastnostmi. Napoved globalnih ali lokalnih nagnjenosti k interakciji z RNA lahko dosežemo s primerjanjem opazovane strukture z znanimi strukturami kompleksov protein-RNA.

Metode, osnovane na strukturah, se lahko ozirajo na uporabljeno metodologijo za določanje proteinske strukture, zato so lahko uporabljene za napovedovanje struktur, pridobljenih z rentgensko kristalografijo, NMR ali teoretičnim modeliranjem. Vsi taki modeli potrebujejo posebno obravnavo. V kristalnih strukturah lahko napovedovanje interakcij na aminokislinah potrebuje spremembe vhodnih podatkov, na primer dodajanje mankajočih neurejenih zank z metodo primerjalnega modeliranja. Poleg tega pa prostorske metode tipično napovejo posamezne modele namesto večih modelov, kar zahteva izbor predstavitvene strukture ali računanje konsenznega modela za NMR podatke. Tudi napovedovanje RNA-vezavnih aminokislin, osnovanih na teoretičnih modelih, zahteva upoštevanje globalnih in lokalnih modelov, ker se napake ter nenatančnosti teoretičnih modelov širijo v napovedane komplekse. Večina proteinov veže RNA kot oligomeri, medtem ko nekatere metode sprejmejo le enostranke verige in monomerne strukture.

2.4.1 Obstoječe rešitve

Različni pristopi za napovedovanje RNA-vezavnih proteinov, osnovanih na strukturni analizi in na prepoznavanju RNA-vezavnih aminokislin, so pregledani v literaturi, vendar so le nekateri algoritmi razviti v splošno namenski obliki, ki je dostopna javnosti. Zlasti število orodij na področju bioinformatike za prostorsko napovedovanje mest aminokislin, ki so v interakciji z RNA, je precej manjše kot metode za napovedovanje DNA in proteinskih vezav. V nadaljevanju bomo opisali najpogostejše metode, ki so dostopne kot spletne strani ali kot samostojen program.

Metode Struct-NB, PRIP, PatchFinderPlus, SPOT in OPRA napovedu-

jejo interakcijo RNA z uporabo lastnosti površine proteina. Na strukturnih značilnostih so uporabljeni klasifikatorji SVM in naivni Bayes. RNABindR metoda kombinira strukturno informacijo z napovedjo hidrofobnosti in entropije na zaporedju. Uspeh metod, ki temeljijo na strukturi, lahko nudi tudi strukturne podrobnosti vezave substratov, vendar je omejen z dostopnostjo kompleksov protein-RNA.

RNABindR je klasifikator, ki napove RNA-vezavna območja [45]. Značilke, uporabljene v tej metodi, so relativna dostopna površina (rASA), entropija zaporedja, hidrofobnost, sekundarna struktura in elektrostatika. Relativna dostopna površina je izračunana s programom Naccess. Entropija zaporedja je ocenjena z uporabo relativne entropije za vsako molekulo iz HSSP podatkovne base. Hidrofobnost vsake aminokisline je pridobljena iz konsenzne normalizirane hidrofobne lestvice avtorjev Sweet in Eisenberg [42]. Poleg tega avtorji uporabijo informacijo in sekundarno strukturo uporabljeno iz baze proteinov. Elektrostatični potenciali so izračunani z uporabo programa APBS. Z uporabo prečnega preverjanja RNABindR prepozna aminokisline s klasifikacijsko točnostjo 85 %.

Struct-NB uporablja zbirko naivnih Bayesovih klasifikatorjev (NBC) in strukturni Gausov naivni Bayesov klasifikator (GNBC) za napovedovanje mest interakcije na proteinu [46]. NBC model je naučen na značilkah zaporedja, ki so enaki kot pri RNABindR algoritmu [45], medtem ko strukturni GNBC uporablja stopnjo nepravilnosti na površini in CX vrednost, ki je definiran kot razmerje med volumnom atomov v 6Å krogli v primerjavi volumnom prazne krogle. Avtorji dosežejo površino pod krivuljo ROC višine 0.75. Analiza pokaže, da so mesta protein-RNA, ki so v interakcij, povezana z višjo stopnjo nepravilnosti na površini v primerjavi z aminokislinami, ki niso v interakciji.

PRIP uporablja NBC in SVM model, kombiniran z matematičnimi grafi lastnosti aminokislin, ki so v interakciji [26]. Aminokislina je klasifici-

rana kot v interakciji glede na lastnosti treh tipov aminokislin v interakciji. Prvi tip je sekvenčno drsno okno velikosti n, drugi je množica n aminokislin, ki so si najbližje v prostoru, tretji tip pa topološki odsek n vozlišč z najmanjšo evklidsko razdaljo do centralnega vozlišča. Metoda za napoved uporabi lastnosti prostorske oblike, dostopne površine, medsebojne centralnosti in zadrževalnega koeficienta. Avtorji dosežejo površino pod krivuljo ROC 0.83.

PatchFinderPlus algoritem uporablja lastnosti proteina in specifične lastnosti, izluščene iz elektrostatičnih odsekov [41]. Avtorji uporabijo SVM za razlikovanje RNA-vezavnih proteinov od pozitivno nabitih proteinov, ki ne vežejo nukleinskih kislin. Metoda je uporabljena na proteinih, ki vsebujejo RNA prepoznavne motive (RRM) in razvrsti RNA-vezavne proteine iz RRM domen, ki so vključeni v protein-protein interakcije. Poleg tega so značilnosti izluščene izmed površinskih odsekov.

SPOT uporabi strukturne poravnave znanih kompleksov protein-RNA in statistično energijsko funkcijo za razločevanje RBP izmed proteinov, ki ne vežejo RNA [60]. Metoda uporabi Z-oceno za merjenje strukturne podobnosti in statistično energijsko funkcijo za merjenje protein-RNA vezavne podobnosti. Prednost te metode je istočasna napoved struktur protein-RNA kompleksov. Metoda dosega MCC oceno 0.72.

OPRA metoda je bila razvita za prepoznavanje mest interakcij protein-RNA na površini proteina [35]. Sprva so avtorji izpeljali napovedno oceno iz verjetnosti interakcije za vsako aminokislino in jo utežili z dostopno površino (ASA). Posamezne verjetnosti se izkažejo kot slab indikator interakcije, zato avtorji uporabijo energijo odsekov, preoblikovano z ocenami sosednjih aminokislin in jih uporabijo za napovedni model. Nato so optimalne energijske ocene izračunali za vsako aminokislino s seštevanjem posameznih ocen sosednjih aminokislin. Metoda pravilno napove 80 % mest na testni množici in nakazuje, da odločilni

dejavniki interakcij protein-RNA ležijo na strani proteina.

KYG uporablja več ocen, ki so osnovane na RNA-vezavnih nagnjenosti posameznih aminokislin in nukleotidov, parov aminokislin in nukelotidov, profilov zaporedij ter njihovih kombinacij [18].

DRNA napove RNA-vezavne proteine in RNA vezavna mesta glede na podobnost z znanimi strukturami. Uporablja strukturno poravnavo z znanimi kompleksi protein-RNA in ocenjevanjem interakcij z DFIRE statistično energijsko funkcijo [58].

2.4.2 Primerjava obstoječih rešitev

V tabeli 2.1 so opisane uspešnosti napovednih metod, ki so jih predstavili avtorji v člankih. Za tiste metode, kjer so bili podatki na voljo, smo MCC vrednosti povzeli iz rezultatov primerjave, opravljene v Puton et. al [36].

Metoda	MCC	AUC	Accuracy
PRINTR		0.83[b]	
PRIP		0.83[b]	
PiRaNhA	0.435[a]	0.822[a]	
KYG	0.382[a]		
DRNA	0.382[a]		
PPRInt	0.339[a]	0.779[b]	
RNABindR	0.317[a]	0.708[a]	0.85[b]
BindN	0.297[a]	0.733[a]	
OPRA	0.296[a]		0.80[c]
Struct-NB		0.75[b]	
PRBR	0.294[a]		

Tabela 2.1: Kvaliteta obstoječih rešitev.

[a]Navedeni so podatki iz študij Puton et. al. [36]
[b]Navedeni so podatki iz študij Cirillo et. al. [6]
[c]Navedeni so podatki iz študij Perez-Cano et. al. [35]

Poglavje 3

Umestitev proteina in RNA

Problem umestitve (angl. docking) proteina in liganda, ki se nanaša na napovedovanje interakcij med makromolekulo, po navadi proteinom in manjšo ciljno molekulo, se pojavi v veliko aplikacijah za prepoznavo molekul, recimo odkrivanje zdravil ter oblikovanje receptorjev in encimov. Problem umestitve je trenutno široko raziskovano področje in je v stopnji hitrega razvoja.

Umestitvene metode se pogosto uporabljajo za napoved 3D struktur makromolekularnih kompleksov. Problem napovedovanja strukture kompleksa lahko razdelimo na dva podproblema. Prvi podproblem je preiskovanje konformacijskega prostora možnih orientacij in pozicij komponent z iskalnim algoritmom. Drugi podproblem pa je razlikovanje ustreznih struktur izmed alternativnih modelov strukturnih kompleksov, ki jih generira iskalni algoritem. Razlikovanje poteka z uporabo ocenjevalne funkcije, ki vodi iskalni postopek in izbere pravilno metodo iz nabora. Mnogo metod združuje obe nalogi, nekatere pa se osredotočijo samo na ocenjevanje kandidatov, in prepustijo generiranje modelov uporabniku.

Idealna umestitvena metoda je zmožna sestaviti strukturo komponent v kompleks in oceniti strukturo, ki je bližje ustrezni strukturi z višjo oceno kot pa neustrezno. V realnosti je struktura kompleksa neznana. Strukture posamezno rešenih vezavnih parov so ponavadi izpostavljene konformacijskim spremembam med povezovanjem v procesu, imenovanem inducirano prilaga-

janje. Umestitveni algoritmi na realnih strukturah morajo dovoljevati take situacije. Korformacijske spremembe so modelirane izrecno z metodami visoke natančnosti, ki naredijo take analize računsko zelo zahtevne ali pa povzročijo določeno stopnjo nejasnosti.

Eden izmed zanimivih aspektov RNA struktur in interakcij protein-RNA je prisotnost posttranskripcijskih sprememb, ki povečajo osnovno množico štirih nukleotidov (A,U,G,C) do več kot 100 variant s spremenjenimi baznimi ali riboznimi deli. Spremenjeni nukleotidi v RNA so odgovorni za mnogo procesov, vključno z RNA pregibanjem in RNA-RNA interakcijami, poleg tega pa tudi specifične protein-RNA prepoznave ter vezave. Spremenjeni nukleotidi so pogosto problematični za razpoložljive umestitvene metode, ker niso prikazani kot tipični potenciali in morajo biti preoblikovani v nespremenjene kandidate v RNA strukturah uporabljenih za umestitev [36].

3.1 Preiskovalne metode

Za napovedovanje realnega načina povezovanja proteina in liganda je tipičen iskalni algoritem, ki vzorči dovolj velik nabor vezavnih možnosti. Njegova naloga je generiranje možnih strukturnih pozicij molekule. Iskalni algoritmi morajo upoštevati stopnjo fleksibilnosti translacije in rotacije liganda, poleg tega pa sodobni umestitveni postopki po navadi obravnavajo ligand kot fleksibilno molekulo. Obstoječi iskalni algoritmi so kategorizirani v tri osnovne tipe: Naključne ali stohastične metode, sistematične in simulacijske metode [24].

Naključne ali stohastične metode vzorčijo razvrstitveni prostor z izvajanjem sprememb liganda v vsakem koraku. Spremembe so nato sprejete ali zavrnjene glede na v naprej določeno verjetnostno funkcijo. Osnovano na naključnih algoritmih je ta metoda nadaljnje klasificirana v tri tipe. Metode z genetski algoritmi vključujejo AutoDock, GOLD in DARWIN [31, 15, 44]. Metode Monte Carlo vključujejo Prodock, ICM, MCDOCK, DockVision in QXP [47, 1, 22, 14, 27]. Metode, ki

uporabljajo Tabu algoritem, spodbujajo učinkovitost s preprečevanjem vračanja na že pregledana stanja. PRO_LEADS [2] je primer metode, ki uporablja Tabu iskalni algoritem.

Sistematične metode kombinatorično preiskujejo konformacijski prostor. Vezi v ligandu se pri vsakem koraku lahko z majhnimi spremembami obračajo za 360 stopinj. Če želimo preprečiti, da izračun postane pretežak zaradi kombinatorične eksplozije, se pogosto problem razdrobi in manjše kose umesti ločeno. Druga strategija je uporaba metod baze podatkov, ki raziskujejo zbirko v predhodno generiranih ligandnih konformacijah. Vsaka razvrstitev v zbirki je lahko obravnavana kot togo telo med umestitvenim procesom. Primeri sistematičnih metod vključujejo DOCK, LUDI, FlexX, ADAM, HammerHead in FLOG [9, 3, 37, 30, 56, 28].

Simulacijske metode ponavadi vključujejo molekularno dinamiko in minimizacijo energije. Te metode imajo slabost, da se ujamejo v lokalnem minimumu in niso tipično uporabljene kot samostoječe iskalne tehnike v dejanski umestitveni nalogi. Namesto več različnih umestitvenih algoritmov uporabljajo simulacijske metode. Ena izmed takih metod je DOCK [9], ki opravlja računanje energijske minimizacije po vsakem koraku sprememb.

Pri pregledovanju struktur za odkrivanje novih zdravil je potrebno pregledati obsežne knjižnice. Pri takih strukturah je učinkovitost pomembnejša od natančnosti. LigandFit in LibDock sta dva primera umestitvenih programov, katerih učinkovitost je prilagojena za virtualni pregled v obsežnih knjižnicah [51, 7]. Vedno je potreben kompromis med učinkovitostjo in natančnostjo za vsak umestitveni algoritem, saj so tiste metode, ki potrebujejo večjo konformacijsko množico in zapletenejšo ocenjevalno funkcijo pogosto natančnejše, a obenem manj učinkovite.

3.2 Ocenjevalne funkcije

Skupaj z iskalnimi metodami se pojavi potreba po ocenjevanju kandidatov generiranih v iskalnem procesu. Uspešna ocenjevalna funkcija mora biti dovolj stroga, da oceni pravilne pozicije z boljšo oceno, istočasno pa ne sme biti računsko prezahtevna. Sedanje ocenjevalne funkcije v široki uporabi so osnovane na polju silnic, na primer CHARMM, AMBER, G-Score in GoldScore [33, 55, 19, 50], in tiste, ki temeljijo na empiričnih podatkih, na primer F-Score, SCORE in X-Score [37, 43, 52].

Z danimi strukturami proteina in liganda je zanesljiva ocena umestitvenega algoritma po navadi povprečna kvadratna razdalja (RMSD) med generirano in eksperimentalno umestitvijo liganda.

Večina protein-RNA umestitvenih metod, ki jih bomo opisali v naslednjem delu tega poglavja, npr. GRAMM, PatchDock in Hex [17, 40, 38] in so zmožne voditi spremenjene nukleotide v RNA molekulah, nimajo primernih ocenjevalnih funkcij za prepoznavo ustreznih struktur RNP kompleksov, zato so potrebne posebne razširitve za ocenjevanje interakcij protein-RNA. V zadnjih nekaj letih, so bile razvite metode statističnih potencialov za ocenjevanje interakcij protein-RNA.

Zheng et. al. [59] so razvili statistični potencial, ki je odvisen od razdalj med vsemi atomi. Dobro deluje na modelih kompleksov protein-RNA, ki so podobni realni strukturi (RMSD <5), vendar med realnim umestitvenim eksperimentom težko doseže komplekse, ki so blizu realni strukturi. V večini primerov vezave protein-RNA se zgodijo zmerne konformacijske spremembe med proteinom in RNA molekulo. V takih primerih so koristne metode z nizko ločljivostjo, ki ignorirajo atomske podrobnosti spremenjene med vezavo.

Drugi potencial so razvili Perez-Cano et. al. in deluje na nivoju nukleotidov. Razvit je bil z namenom izboljšave FTDock potenciala in ni dostopen kot samostojen program [34].

3.3 Obstoječe metode za modeliranje kompleksov protein-RNA

Dosedaj je bilo razvitih veliko protein-protein umestitvenih metod, medtem ko je število metod za modeliranje kompleksov protein-RNA še vedno omejeno. V tem poglavju bomo opisali nekatere javno dostopne spletne aplikacije in samostojne umestitvene programe, ki za vhodne podatke sprejmejo tako protein kot RNA koordinate ter ocenjevalne funkcije za izbor najbolj natančnih modelov izmed množice kandidatov. Za komplekse protein-RNA je bilo razvitih zelo malo metod. Namesto tega obstaja mnogo metod za modeliranje protein-protein kompleksov, ki so bile prilagojene, da delujejo tudi za RNA.

H-DOCK sprva uporablja deli in vladaj strategijo za razvrščanje intermolekularnih načinov med proteini in ligandi z ujemanjem vodovikovih vezi [24]. Vsaka umestitev liganda je izračunana glede na ustrezno geometrijo vodikovih vezi in uporablja ocenjevalno funkcijo. Ocenjevalna funkcija po večini odraža van der Waals interakcije, ki so uporabljene za ocenjevanje umestitve liganda. H-DOCK so avtorji testirali za togo in prožno ligandno umestitev. Prožna umestitev je implementirana s ponavljanjem togih umestitev različnih konformacij manjših molekul in njihovo razvrstitvijo.

H-DOCK so avtorji testirali za toge ligande na množici 271 kompleksov, kjer je vsaj ena intermolekularna vodikova vez. H-DOCK na tej množici doseže uspešnost 91.1 %. Za prožne ligande je bil H-DOCK testiran na drugem naboru 93 kompleksov, ki vsebuje realne ligandne razvrstitve kot tudi 100 kandidatov tvorjenih z orodjem AutoDock [31]. H-DOCK je dosegel uspešnost 81.7 %.

H-DOCK se lahko potencialno uporablja za obsežno virtualno pregledovanje kot predfilter za natančnejše, vendar manj učinkovite umestitvene algoritme.

QUASI-RNP in DARS-RNP Tuszynska in Bujnicki sta razvila dva potenciala s srednjo ločljivostjo za ocenjevanje modelov RNP kompleksov [49]. Prva uporablja navidezni kemični potencial (QUASI-RNP), druga pa uporablja kandidate za referenčno stanje potenciala (DARS-RNP). Ti potenciali so osnovani na poenostavljeni predstavitvi proteinov in RNA uporabijo enako matematično bazo.

DARS-RNP in QUASI-RNP programa imata tudi funkcijo za razvrščanje najbolje ocenjenih struktur. To pomaga pri prepoznavanju podobnih struktur z dobrimi ocenami, ki z višjo verjetnostjo predstavljajo realne konformacije.

Haddock uporablja biokemične in biofizične podatke o interakcijah kot omejitve. Omogoča umestitev različnih molekul, med drugim proteinov, nukleinskih kislin in majhnih molekul. Dostopen je kot samostojen program in spletni strežnik [8].

GRAMM je program za umestitev z nizko natančnostjo. Opravlja šestdimenzionalno iskanje skozi translacije togih teles in rotacije molekule liganda. Ne dovoljuje uporabe omejitev med postopkom umestitve. Zmožen je generiranja kandidatov za vsako molekulo, vendar zahteva posebno zunanjo ocenjevalno funkcijo za komplekse, ki ne vsebujejo proteinov [17].

Hex omogoča umestitev protein-protein in protein-nukleinske kisline. Uporablja sferično polarno Fourierjevo korelacijo (SPF). Znanje vezavnih mest je lahko uporabljeno za optimizacijo računanja, ocenjevalna funkcija pa vsebuje ujemanje oblike in elektrostatiko. Metoda nima posebne funkcije za komplekse protein-RNA [38].

PatchDock je molekularni umestitveni algoritem osnovan na geometriji [40]. Razvit je bil za napovedovanje kompleksov med dvema proteinoma in proteinom ter majhno molekulo. Generira lahko položaje za protein-nukleinska kislina komplekse, ampak nima ocenjevalne funkcije za iden-

tifikacijo dobrih modelov. Dovoljuje definicijo potencialnih vezavnih mest v ligandu in receptorju. Dostopen kot samostojen program in kot spletni strežnik.

FTDock opravlja umestitev togega telesa. Program je bil razvit za protein-protein umestitve, vendar sprejme tudi RNA in DNA molekule. Nima specializirane ocenjevalne funkcije za komplekse protein-RNA [13].

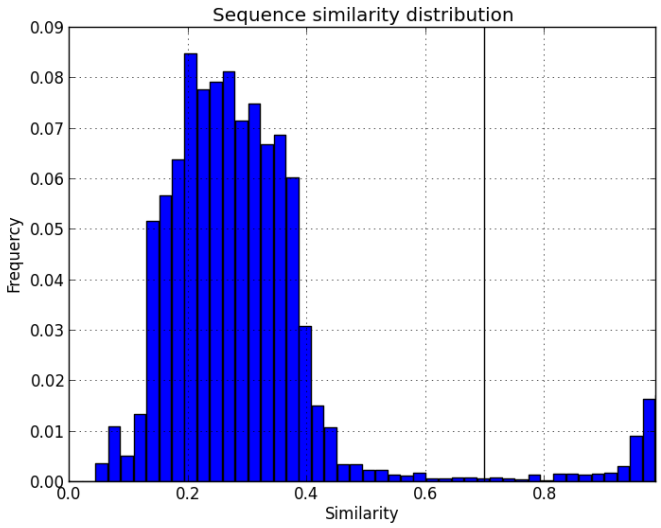
Poglavje 4

Napovedovanje mest interakcije

V tem poglavju opišemo postopek modeliranja mest interakcije na podlagi več kot 200 neredundančnih podatkovnih struktur. Najprej predstavimo nabor struktur, uporabljenih pri analizi in način določanja mest interakcij. V nadaljevanju poglavja podrobneje opišemo uporabljene značilke na proteinu, značilke na RNA in značilke interakcij. Zgradimo več modelov, ki uporabljajo različne nabore značilk in različne klasifikatorje. Sledi opis gradnje napovednih modelov in način generiranja podatkov, osnovanih na proteinski verigi ter na verigi RNA. Na koncu opišemo še merila za vrednotenje napovednih modelov.

4.1 Podatki

Skupno smo analizirali 822 struktur protein-RNA iz protein data bank (PDB) zbirke [61]. Izbrali smo strukture, ki vsebujejo vsaj eno RNA molekulo. Uporabili smo najboljše podatke, pridobljene z rentgetsko kristalografijo z natančnostjo boljšo od 4 Å in podatke pridobljene z NMR metodo. Izmed teh smo odstranili komplekse, ki vsebujejo ribosomalne proteine in proteinske strukture z več kot 500 aminokislinami. V nasprotnem primeru je število učnih primerov, dobljenih iz teh struktur, preveliko in postane informacija pridobljena iz enostavnejših kompleksov zanemarljivo majhna. Izmed pro-

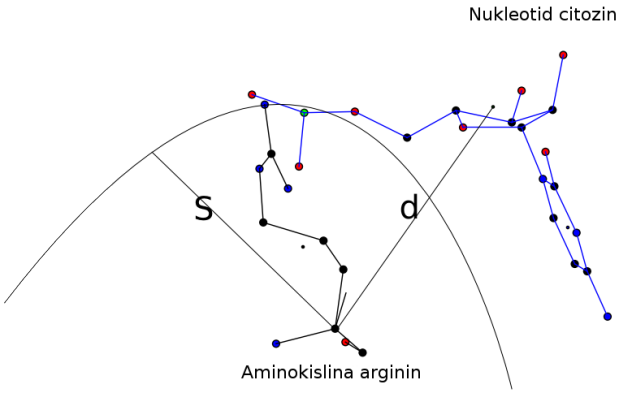


Slika 4.1: Porazdelitev podobnosti med pari pozameznih verig.

stalih kompleksov smo odstranili tiste, ki imajo manj kot pet nukleotidov na verigi RNA, ker iz njih ne moremo izluščiti vseh potrebnih značilk, kot so na primer koti med sosednjimi nukleotidi in soseščina nukleotidov. Preostane nam 360 struktur protein-RNA, iz katerih potem odstranimo še podvojene strukture in strukture, kjer je podobnost zaporedja večja od 70 %. Histogram porazdelitve podobnosti parov verig je prikazan na sliki 4.1. Navpična črta prikazuje mejo 70 %, ki je taka zato, ker je v območju 60 %–80 % zelo malo verig in dobro razdeli podatke na dve ločeni skupini. Ostane nam 208 struktur protein-RNA, ki smo jih uporabili v modelu za napovedovanje mest interakcije in pri optimizacijskem algoritmu, ki je opisan v poglavju 5.

4.2 Mesta interakcij

Pri določanju parov aminokisline in nukleotida, smo med njima izmerili razdaljo. Pare, kjer je relativna razdalja manjša od 4 Å, smo označili, da so v interakciji. Relativna razdalja med aminokislino in nukleotidom je definirana



Slika 4.2: Razdalja med aminokislino in nukleotidom je označena z d, povprečna dolžina aminokisline je označena s S.

kot razdalja med atomom $C\alpha$ aminokisline in centrom riboze nukleotida (d), kateri odštejemo povprečno dolžino aminokisline (S). Dolžina aminokisline S je definirana kot razdalja od $C\alpha$ do najbolj oddaljenega atoma na aminokislini in se povpreči čez vse atome, zaradi zmanjšanja vpliva merskih napak v 3D koordinatah posameznih atomov. Shema računanja relativne razdalje je prikazana na sliki 4.2, kjer je S povprečna dolžina aminokisline in d razdalja med $C\alpha$ in centrom riboze. Relativna razdalja je definirana kot vrednost $d - S$.

4.3 Značilke

V tem poglavju bomo natančno opisali značilke, uporabljene v modelu. Ločili smo jih v tri logične skupine. Prve so značilke na proteinu. To so tiste, ki jih lahko dobimo, če imamo na voljo izključno strukturo proteina, podatkov o verigi RNA pa ni na voljo. Druga skupina so značilke na podlagi strukture RNA, ki so na voljo tudi brez podatkov o proteinu. Značilke interakcij pa so tiste značilke, ki za izračun potrebujejo verigo proteina in verigo RNA.

4.3.1 Značilke na proteinu

Protein je molekula, ki jo sestavljajo manjši gradniki imenovani aminokisline, ki se med seboj povezujejo s peptidno vezjo. Temelj proteina v vsaki aminokislini predstavlja sekvenca atomov N-$C\alpha$-C. Osnovna veriga je precej prilagodljiva in omogoča fleksibilno obliko proteina. Vsaka aminokislina ima drugačno zaporedje atomov in povzroči specifične fizikalne ter kemijske lastnosti. Obstaja 20 vrst aminokislin najdenih v evkariontskih organizmih. Različne aminokisline imajo različne nagnjenosti k interakciji in različne 3D lokalne strukturne značilnosti.

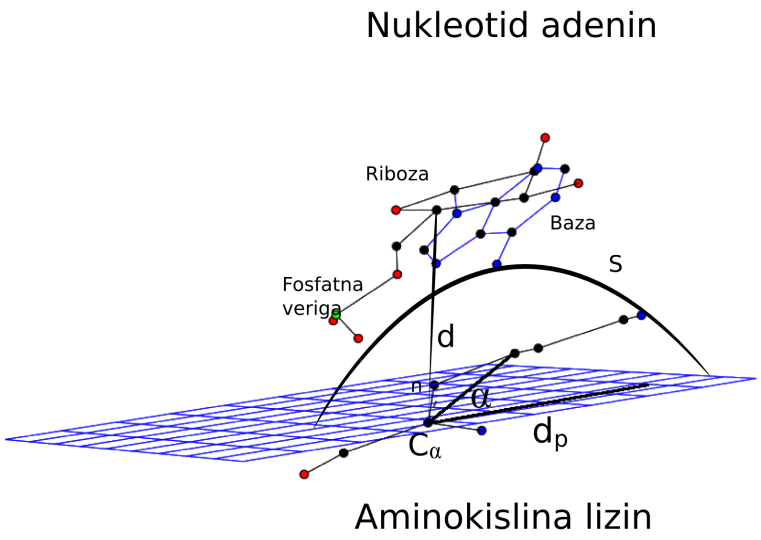
Vrsta aminokisline predstavlja eno izmed 20 aminokislin.

Smer aminokisline je določena kot razlika med koordinatami atoma $C\alpha$ in koordinato točke, ki je na sredini med atomoma N in C v temelju iste aminokisline. Temelj aminokisline je del osnovne verige proteina.

Dolžina aminokisline predstavlja povprečno dolžino aminokisline. Dolžino aminokisline dobimo z računanjem razdalje med atomom $C\alpha$ in atomom, ki je najbolj oddaljen od $C\alpha$. Nato izračunamo povprečje teh razdalj za vsako aminokislino posebej. Dolžina aminokisline je prikazana na sliki 4.3 z oznako S.

Dostopna površina aminokisline je odvisna od vrste aminokisline. Dostopna površina molekule je definirana kot sled centra sfere z radijem vodne molekule, ki se giblje okrog površine modela molekule. Lestvica je povzeta iz Miller et. al. [29]. Vrednosti dostopne površine, ki smo jih uporabili, so prikazane v tabeli 4.1.

Polarnost aminokisline je fiziokemična lastnost, ki je odvisna od vrste aminokisline. Polarna lestvica, prikazana v tabeli 4.1, je osnovana na povprečni razvrstitvi aminokislin v 38 objavljenih hidrofobnih lestvicah [48].



Slika 4.3: Razdalja med aminokislino in nukleotidom je označena z d, dolžina aminokisline je označena s S, dolžina projekcije aminokisline je označena z d_p, kot med aminokislino je označen z α.

Dolžina projekcije aminokisline izračunamo tako, da izračunamo koordinate centra vseh atomov v aminokislini X_c. Nato izračunamo ravnino, ki je določena z vektorjem $C\alpha$ in smerjo aminokisline. Koordinate centra aminokisline X_c nato projeciramo na to ravino v točko X_c' in izračunamo razdaljo med točko $C\alpha$ in točko X_c'. Aminokislinam se nagibi spreminjajo ob interakciji, kar se pozna pri projekciji na ravnino. Bolj nagnjene aminokisline imajo daljšo projekcijo. Na sliki 4.3 je označena dolžina projekcije d_p na ravnino, ki je definirana z vektorjem smeri aminokisline n in točko $C\alpha$.

Relativna dolžina projekcije je definirana kot dolžina projekcije deljena z dolžino aminokisline. Namen te značilke je zmanjšati vpliv dolžine

aminokislin pri dolžini projekcije, da je lažje razvidno pri katerih se aminokislina bolj upogne.

Kot aminokisline glede na projekcijo predstavlja kot pod aminokislino glede na ravnino, ki jo določa smer aminokisline. Izračunamo ga tako, da najprej izračunamo projekcijo centra aminokisline na ravnino, ki je določena s $C\alpha$ in smerjo aminokisline podobno, kot smo izračunali dolžino projekcije. Nato izračunamo kot med vektorjema $X'_c - C\alpha$ in $X_c - C\alpha$, kjer točka X_c predstavlja center vseh atomov v aminokislini, točka X'_c pa predstavlja projecirano točko centra aminokisline. Ta značilka pove pod kakšnim kotom je aminokislina glede na svoj smerni vektor in omogoča, da vidimo spremembe kota aminokisline pri interakcijah. Kot aminokisline glede na projekcijo je prikazan na sliki 4.3 s kotom α.

Število aminokislin v krogli določimo tako, da za vsako aminokislino v proteinu izračunamo skupno število aminokislin, ki so oddaljene za največ r. Računali smo več takih značilk z različnimi polmeri krogle.

Število aminokislin v zgornji polkrogli določimo tako, da za vsako aminokislino (A_0) v proteinu najprej dobimo druge aminokisline (A_1, ..., A_n), ki so oddaljene za največ r. Nato izmed aminokislin (A_1, ..., A_n) preštejemo vse, za katere velja, da je na zgornji strani ravnine, določene z atomom $C\alpha$ in smernim vektorjem aminokisline A_0. Računali smo več takih značilk z različnimi vrednostmi polmera.

Aminokislina	Povprečna dolžina (Å)	ASA	Polarnost
A	2.40	67	9
C	2.80	104	7
D	3.57	103	19
E	4.54	138	18
F	5.13	175	2
G	2.39	0	11
H	4.58	151	10
I	3.70	140	1
K	5.68	167	20
L	3.88	137	3
M	4.58	160	5
N	3.59	113	16
P	2.45	105	13
Q	4.53	144	17
R	6.47	196	15
S	2.44	80	14
T	2.54	102	12
W	6.09	217	6
V	2.54	187	8
Y	6.45	117	4

Tabela 4.1: Povprečna dolžina, dostopna površina in polarnost aminokislin.

4.3.2 Značilke na RNA

RNA je sestavljena iz štirih različnih nukleotidov. Te značilke so uporabne za algoritem umestitve, kjer potrebujemo podatke o strukturnih značilnostih verige RNA.

Vrsta nukleotida predstavlja enega izmed (A,C,G,U) nukleotidov.

Dolžina nukleotida je razdalja med centrom baze in centrom riboze istega nukleotida.

Razdalja med sosednjimi nukleotidi predstavlja razdaljo med trenutnim nukleotidom N_n in naslednjim nukleotidom N_{n+1}, če ta obstaja.

Kot med sosednjimi nukleotidi predstavlja kot med vektorjem do predhodnega nukleotida $N_{n-1} - N_n$ in vektorjem do naslednjega nukleotida $N_{n+1} - N_n$. Dobimo ga s skalarnim produktom vektorjev $(N_{n-1} - N_n) \cdot (N_{n+1} - N_n)$.

4.3.3 Značilke interakcije

Za napovedovanje s strojnim učenjem potrebujemo razdaljo med aminokislino in najbližjim nukleotidom, zato da določimo, ali je pri danem paru prišlo do interakcije ali ne. Ostale značilke so uporabne za napovedni model, kjer imamo prisotni obe strukturi in želimo najti značilke, ki najbolj izboljšajo točnost modela.

Razdalja med aminokislino in nukleotidom je razdalja med atomom $C\alpha$ proteina in nukleotidom, kjer odštejemo povprečno dolžino aminokisline v interakciji.

Kot interakcije je kot med smernim vektorjem aminokisline in vektorjem, ki je definiran kot razlika med centrom riboze nukleotida in atomom $C\alpha$ na aminokislini.

Stran proteinske verige pove, ali je RNA na isti strani glavne verige proteina kot je aminokislina (angl. sidechain), ali na hrbtni strani aminokisline (angl. backbone). Če je nukleotid na isti strani verige kot aminokislina, je vrednost značilke 1, v nasprotnem primeru je vrednost značilke -1.

Stran verige RNA pove, ali se RNA približa aminokislini s hrbtno stranjo verige ali z bazno stranjo verige. Nukleotidi v interakciji imajo večjo verjetnost, da bo riboza bližje aminokislini kot pa baza. Hrbtna stran RNA je negativno nabita in se posledično obrne proti proteinu, ki je pozitivno nabit. Vrednost značilke je 1, kadar je riboza bližje aminokislini, če je baza bližje aminokislini je vrednost značilke -1.

Razdalja do riboze je določena kot dolžina vektorja med atomom $C\alpha$ in centrom riboze.

Razdalja do baze je določena kot dolžina vektorja med atomom $C\alpha$ in centrom dušikove baze.

Kot do riboze je definiran s kotom med vektorjem razdalje $C\alpha$ do centra riboze in smernim vektorjem aminokisline.

Kot do baze je podoben kotu do riboze, definiran s kotom med vektorjem $C\alpha$ do centra dušikove baze in smernim vektorjem aminokisline.

Število interakcij v bližini izračunamo tako, da preštejemo vse pare aminokislin z nukleotidom, ki so v interakciji in so bližje od dane razdalje r. Izhajamo iz predpostavke, da je pri aminokislinah v interakciji večja verjetnost, da bodo sosednje aminokisline prav tako v interakciji. Pri tej značilki smo uporabili več različnih vrednosti r.

4.4 Gradnja modelov

V našem postopku najprej preberemo vhodne podatke, iz njih izluščimo geometrijske lastnosti in pretvorimo v obliko, ki vsebuje verigo proteina z lokalnimi strukturnimi lastnostmi posameznih aminokislin in verigo RNA z lokalnimi strukturnimi lastnostmi nukleotidov. Nato te podatke preberemo in izračunamo lokalne 3D lastnosti interakcij po dveh različnih načinih pridobivanja podatkov. Pristopa sta osnovana na različnih referenčnih verigah, prvi pristop uporablja proteinsko verigo, drugi pa verigo RNA. Po pridobivanju interakcij na podlagi referenčne verige se seznam interakcij zakodira v matrično obliko. Na teh podatkih se izvede metode strojnega učenja in shrani stanje modela, kar se potem uporabi v umestitvenem algoritmu. V nadaljevanju bomo opisali pridobivanje podatkov glede na referenčno verigo proteina in verigo RNA.

1. V pristopu proteinske referenčne verige za vsako aminokislino na proteinu poiščemo najbližji nukleotid v prostoru in izračunamo relacije med njima. Ta množica podatkov ima za vsako strukturo natančno toliko primerov kot je aminokislin na proteinu, vendar so nekateri nukleotidi uporabljeni večkrat, nekateri pa nikoli. Algoritem pridobivanja podatkov na podlagi proteinske referenčne verige je opisan v psevdokodi na sliki 4.4.

```
function proteinReferenceChain(protein, rna)
pairs ← list()
for aminoacid in protein do
  nucleotide ← closestNucleotide(aminoacid, protein, rna)
  pairs ← pairs + interaction(aminoacid, nucleotide)
end for
return pairs
```

Slika 4.4: Postopek gradnje modela s pristopom proteinske refenčne verige.

2. Pri referenčni verigi RNA za vsak nukleotid na verigi RNA poiščemo najbližjo aminokislino v prostoru in izračunamo relacije med njima. Ta množica vsebuje toliko primerov, kolikor je nukleotidov na verigi RNA, vendar se veliko aminokislin uporabi večkrat, veliko pa se jih ne uporabi. Algoritem pridobivanja podatkov na osnovi referenčne verige RNA je opisan s psevdokodo na sliki 4.5.

```
function rnaReferenceChain(rna, protein):
pairs ← list()
for nucleotide in rna do
  aminoacid ← closestAminoacid(nucleotide, protein, rna)
  pairs ← pairs + interaction(aminoacid, nucleotide)
end for
return pairs
```

Slika 4.5: Postopek gradnje modela s pristopom refenčne verige RNA.

4.4.1 Vrednotenje napovednega modela

Pri klasifikaciji interakcij smo uporabili klasifikacijska drevesa, naivni Bayesov klasifikator (NB) naključne gozdove (RF) in SVM. Z metodo prečnega preverjanja smo napovedali interakcije in izračunali vrednosti metrik, ki jih predstavljajo enačbe 4.1, 4.2, 4.3 in 4.4. Poleg tega smo izračunali površino pod krivuljo ROC (AUC) za vsako izmed metod strojnega učenja. Ko te metrike izračunamo še na večinskem klasifikatorju, jih lahko primerjamo z metodami strojnega učenja in tako ovrednotimo naš model.

Preciznost (angl. precision) pove, koliko izmed pozitivno napovedanih primerov je res v interakciji. Preciznost je definirana z enačbo 4.1. Priklic (angl. recall), ki je definiran z enačbo 4.2, pove število mest v interakciji, ki smo jih pravilno napovedali izmed vseh mest interakcij. Klasifikacijska točnost (angl. accuracy) predstavlja število vseh pravilnih zadetkov v primerjavi z vsemi napovedmi. Klasifikacijska točnost je definirana z enačbo 4.3. Enačba 4.4 predstavlja korelacijski koeficient Matthew, ki vrača vrednosti

med -1 in 1, 0 pa predstavlja napoved, ki je primerljiva z naključno.

TP v enačbah 4.1, 4.2, 4.3 in 4.4 predstavlja število pravilno napovedanih primerov, ki pripadajo pozitivnemu razredu. FP predstavlja število primerov negativnega razreda, ki smo jih napovedali pozitivno. FN predstavlja število primerov pozitivnega razreda, ki smo jih napovedali negativno, TN pa predstavlja pravilno napovedane primere, ki so v negativnem razredu.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.3}$$

$$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.4}$$

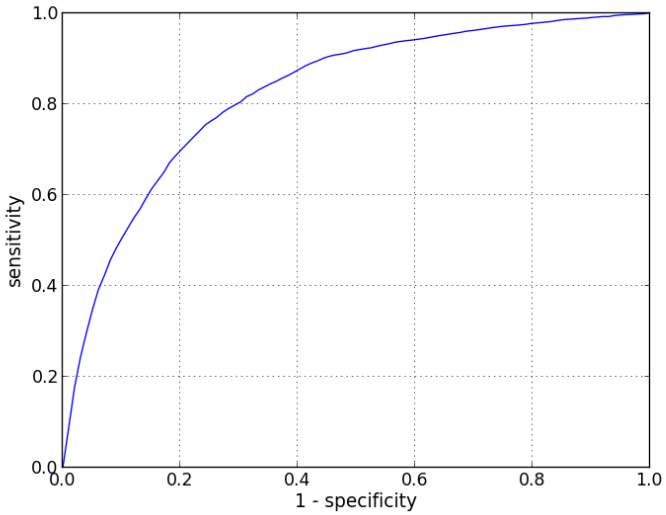
Senzitivnost (angl. sensitivity) predstavlja relativno število pravilno klasificiranih pozitivnih primerov (enačba 4.5). Specifičnost (angl. specificity) predstavlja relativno število pravilno klasificiranih negativnih primerov (enačba 4.6).

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4.5}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4.6}$$

Površina pod krivuljo ROC

Krivulja ROC je graf, ki pokaže kvaliteto binarnega klasifikatorja. Primer krivulje ROC je prikazan na sliki 4.6. Na x osi je prikazano relativno število napačno klasificiranih negativnih primerov (1 - specifičnost). Na y osi pa je prikazano relativno število pravilno klasificiranih pozitivnih primerov (senzitivnost). Bližje kot je krivulja zgornjemu levemu kotu, tem boljši je klasifikator.



Slika 4.6: Primer tipične krivulje ROC

Poglavje 5

Umestitev lokalne strukture protein-RNA

V okviru tega poglavja predlagamo in podrobno opišemo metodo za umestitev lokalne strukture molekule RNA, dolžine do pet mest. Metoda uporablja statistično porazdelitev značilk, opisanih v poglavju 4, kot ocenjevalno funkcijo za razlikovanje med različnimi kandidati RNA pozicij. Iskanje rešitve poteka s preiskovalnim algoritmom, ki uporablja metodo Monte Carlo za ponavljanje iterativnih optimizacij. Iterativna optimizacija preiskuje prostor s sistematičnim generiranjem kandidatov, ki jih izbere glede na najboljšo vrednost ocenjevalne funkcije. Za testiranje umestitvenega algoritma smo iz struktur izluščili zaporedje nukleotidov verige RNA in 3D strukturo proteina. Umestitveni algoritem smo vrednotili z evklidsko razdaljo med 3D strukturo RNA, pridobljeno iz podatkov in rešitvijo algoritma. Algoritmu smo priredili več uteži, ki določajo vpliv posameznih značilk. Testirali smo kvaliteto različnega nabora uteži, na koncu pa smo preizkusili še metodo naključnega vzorčenja in jo primerjali z optimizacijskim algoritmom.

5.1 Preiskovalni algoritem

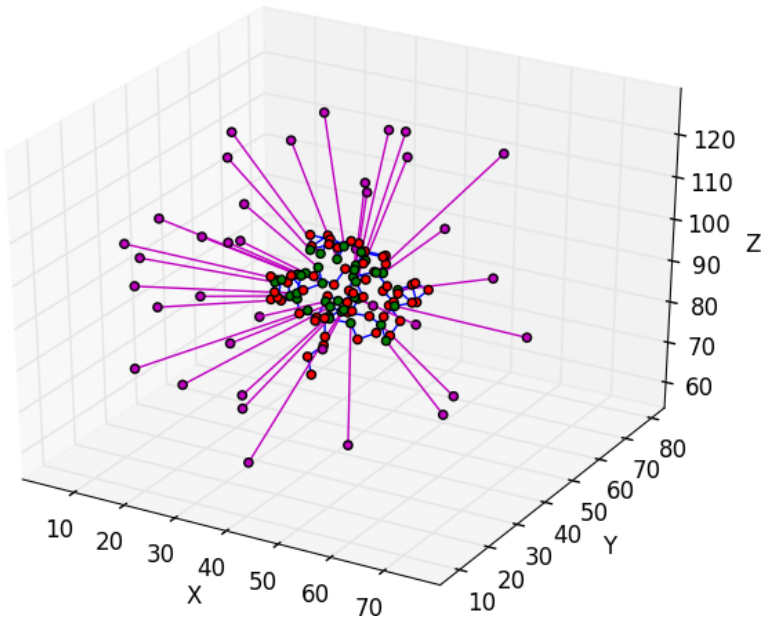
Optimizacijski postopek, implementiran v okviru tega dela, je sestavljen iz več stopenj. Na začetku poteka generiranje začetnih položajev, ki so sestavljeni iz točk v okolici proteina. V naslednjem koraku za vsako začetno točko izvedemo umestitveni algoritem, ki je sestavljen iz preiskovalnega algoritma in ocenjevalne funkcije. Ocenjevalna funkcija v vsakem koraku oceni in izbere boljšo rešitev glede na statistično funkcijo, ki uporablja verjetnostne matrike. Verjetnostne matrike vsebujejo frekvence različnih kombinacij značilk, ki jih dobimo z analizo množice podatkov o 208 strukturah protein-RNA.

5.1.1 Generiranje začetnih položajev

Položaji so generirani tako, da nabor izbranih točk projeciramo na kroglo, ki obkroža celoten protein. Krogla mora biti dovolj majhna, da razdalja do kandidatov ni večja od realnih razdalj do nukleotidov, ki jih pridobimo iz podatkovnih struktur. V našem primeru smo za polmer krogle vzeli polovico maksimalne razdalje med aminokislinami v proteinu, tako da je v notranjosti krogle cel protein. Če točk ne bi projecirali na kroglo, bi prišlo do situacij, kjer je RNA znotraj strukture proteina.

V kodi na sliki 5.2 funkcija *predictInitialPositions* izračuna projekcije začetnih točk na kroglo. Začetne točke predstavljajo mesta interakcij, ki jih dobimo z napovednim modelom, opisanem v poglavju 4. Tem točkam dodamo še nekaj naključno generiranih točk, da vključimo še območja, ki jih napovedni model ni zaznal in za tiste primere, kjer napovedni model nima pozitivih predikcij.

Za vsako analizirano strukturo smo generirali 30 točk in algoritem ponovno pognali na vsaki izmed njih. Nekaj izmed teh začetnih točk smo generirali naključno, odvisno od tega, koliko točk na začetku napove model. Če je bilo število pozitivnih napovedi veliko, smo izmed njih izbrali le naključni vzorec. Primer take krogle, generirane za kompleks protein-RNA z zapisom v PDB 1sj3, je prikazan na sliki 5.1. Zelene točke so aminokisline, kjer model



Slika 5.1: Pozitivno in negativno napovedana mesta ter koordinate začetnih položajev optimizacije na proteinu z zapisom v PDB 1sj3.

napove interakcijo, rdeče točke so aminokisline, kjer model napove odsotnost interakcije, z vijolično zo označene projecirane točke na kroglo. Vijolične točke so začetni položaji različnih iteracij metode Monte Carlo.

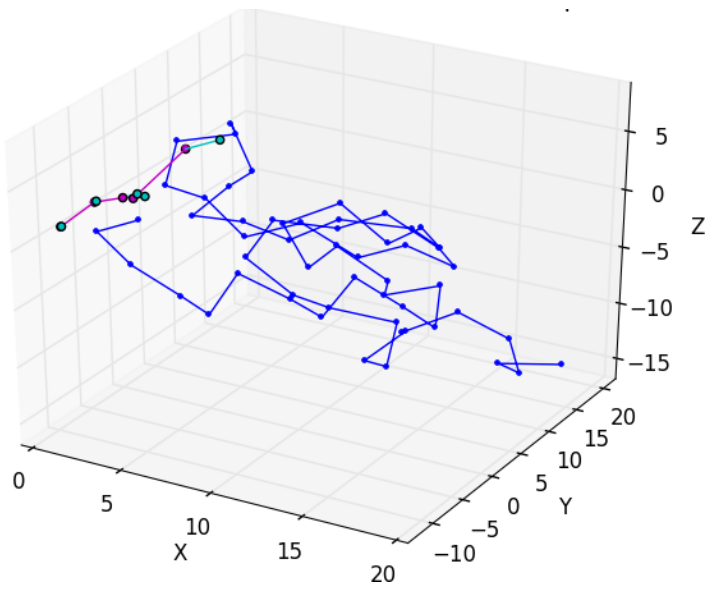
5.1.2 Metoda Monte Carlo

V veliko primerih se optimizacija preiskovalnega algoritma ustavi v lokalnem minimumu zaradi nihanj v statistični funkciji posameznih značilk. Razlog za to je zvijanje zaporedja RNA v manj ugodno lego, včasih pa je že začetna točka v neugodni legi. Temu se poskusimo izogniti z uporabo metode Monte Carlo, ki algoritem začne od začetka iz druge točke vsakič, ko zazna, da je prišlo do konvergence. Omejitev za ustavitev iteracije metode Monte Carlo je nespremenjena ocena preiskovalnega algoritma tri zaporedne iteracije. Potek algoritma je predstavljen s psevdokodo na sliki 5.2.

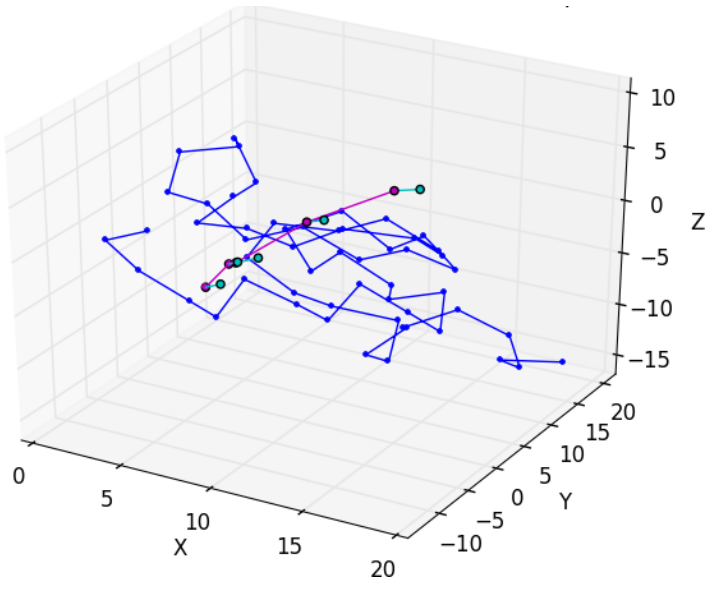
```
function docking():
solutions ← list()
initialPositions ← predictInitialPositions()
for position in initialPositions do
  score ← 0
  while not stoppingCondition() do
    candidateList ← generateCandidates(position)
    for candidate in candidateList do
      if evaluate(candidate) > score then
        score ← evaluate(candidate)
        position ← candidate
      end if
    end for
  end while
  solutions ← solutions + position
end for
return solutions
```

Slika 5.2: Algoritem za umeščanje RNA na protein.

Na začetku vsake iteracije se na novo generira sekvenca 10 točk, ki predstavljajo pet zaporednih nukleotidov, vsak označen z dvema točkama. Prva je center riboze in druga je center baze nukleotida. Preiskovalni algoritem v vsakem koraku generira nove možne kandidate strukture RNA. Nato vse možne kandidate oceni glede na ocenjevalno funkcijo. Če obstaja kandidat, ki je ocenjen boljše od predhodne ocene strukture RNA, potem odsek z najboljšo oceno zamenja predhoden odsek. Iteracija se ponavlja, dokler ni tri zaporedne poteze ocena ista. Nato se postopek ponovi z novo začetno točko. Primera začetnega in končnega stanja preiskovalnega postopka sta prikazana na slikah 5.3 in 5.4. Z modro je označena struktura proteina, z vijolično pa predlagana lokalna struktura RNA.



Slika 5.3: Začetna pozicija iteracije algoritma za protein z zapisom PDB 1a1t.



Slika 5.4: Končna pozicija iteracije algoritma za protein z zapisom PDB 1a1t.

5.1.3 Generiranje kandidatov

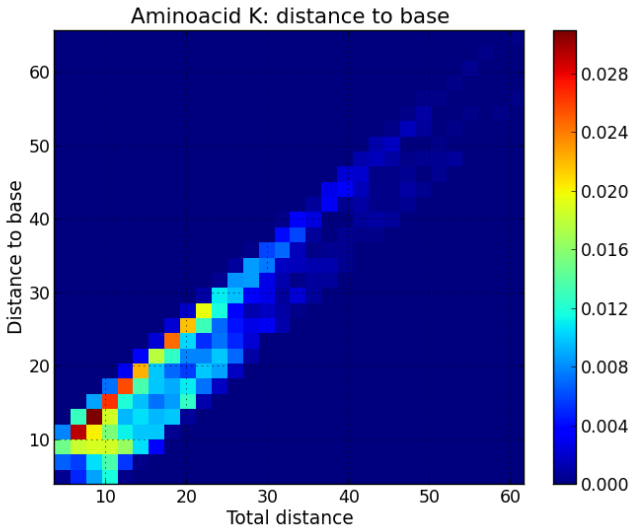
Algoritem v vsakem koraku generira nove kandidate. V poglavju 3 smo opisali osnovne tipe preiskovalnih metod, ki so naključna, sistematična in simulacijska metoda. Kandidate lahko generiramo naključno, vendar je tak proces prepočasen, da bi uspešno preiskali kombinatorični prostor. Nekaj naključnih položajev vseeno dodamo, da proces lažje stopi iz lokalnega minimuma, kadar ostali premiki ne dajejo dobrih rezultatov. Naša metoda sistematično preiskuje prostor in v vsaki iteraciji izračuna pet najbolj tipičnih premikov.

- Premik nukleotida bližje k proteinu – najbolj pogosta uporabljena poteza, ki proteinu približa verigo RNA.
- Premik nukleotida stran od proteina – umik nazaj, kadar pride RNA preblizu proteina.
- Premik nukleotida k levemu sosedu – za uravnavanje razdalj med sosednjimi nukleotidi.
- Premik nukleotida k desnemu sosedu – za uravnavanje razdalj med sosednjimi nukleotidi.
- Premik baze glede na ribozo – ne premakne hrbtne verige, služi za popravljanje dolžine nukleotida, brez da poslabša pozicijo verige.

5.2 Verjetnostne matrike

Verjetnostne matrike predstavljajo 2D histogram, ki opisuje relacije med vrednostjo značilke in razdaljo do najbližje aminokisline. Vsaka aminokislina ima drugo verjetnostno matriko za določeno značilko, saj se porazdelitve značilk med aminokislinami močno razlikujejo. S pomočjo matrik, za neki nukleotid, najdemo najbližjo aminokislino in upoštevamo, s kakšno verjetnostjo se kombinacija vrednosti značilke ter razdalje pojavi v realnih strukturah. Primer take matrike je predstavljen na sliki 5.5 in sliki 5.6, kjer bolj rdeče

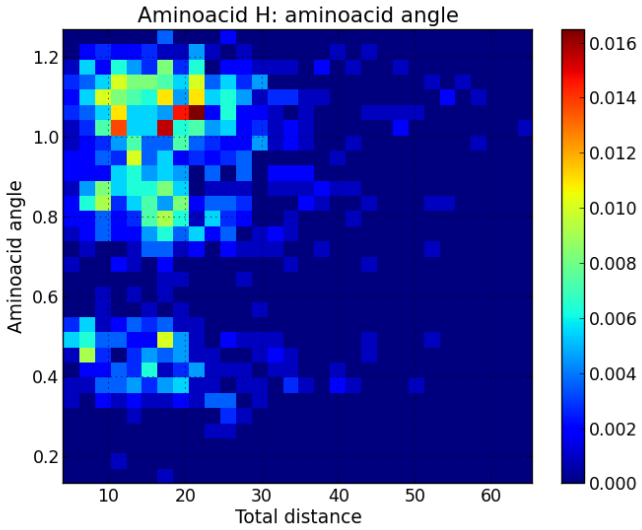
barve predstavljajo višjo verjetnost pojavitve. Te verjetnosti se nato vsaka s svojo utežjo upoštevajo v ocenjevalni funkciji.



Slika 5.5: Porazdelitev razdalje med aminokislino in bazo glede na oddaljenost do strani verige, na kateri so baze.

Matrike favorizirajo določene vrednosti značilk pri dani razdalji in tako vodijo optimizacijski postopek k verjetnejšim situacijam. Nekatere verjetnostne matrike so odvisne od značilke proteina, ki ga med umestitvijo ne spreminjamo. V primeru, da ne spada med tipično pojavljajoče oblike, na primer kot aminokisline velikosti 0.6 za aminokislino histidin (slika 5.6), potem optimizacijski postopek dobiva slabše ocene v bližini te aminokisline. Posledično lahko zato ne skonvergira do proteina. Ker neskonvergirane rešitve na koncu odstranimo, takšna neugodna mesta na proteinu dobijo manj kandidatov in dajo prednost ugodnejšim pozicijam na proteinu.

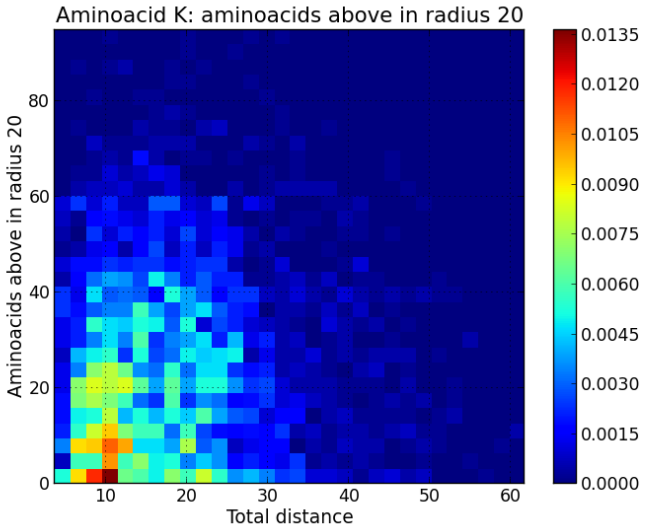
Slike 5.7, 5.9 in 5.8 predstavljajo verjetnostne matrike značilke, ki opisuje število aminokislin v zgornji polkrogli nad aminokislino za tri različne aminokisline: lizin, glutamična kislina in levcin. Lizin je predstavnik pozitivno nabitih aminokislin, glutamična kislina negativno nabitih aminokislin, leucin pa spada med hidrofobne aminokisline. Na osi x je prikazana razdalja med



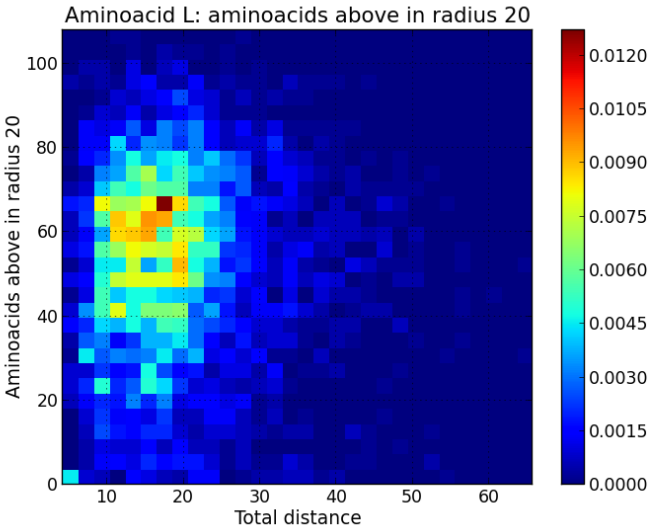
Slika 5.6: Porazdelitev kota pod aminokislino in njeno projekcijo glede na oddaljenost do verige RNA.

parom aminokisline in nukleotida, na osi y pa število drugih aminokislin, ki so nad to aminokislino. Opazno je, da imajo te aminokisline različne porazdelitve razdalj in vrednosti značilke. Lizin ima v večini primerov manjše število okoliških aminokislin, saj je velikokrat v interakciji in je zato potreben prostor nad aminokislino. Glutamična kislina ima večjo verjetnost, da bo število okoliških aminokislin manjše, vendar s to razliko, da je večina nukleotidov oddaljena za vsaj 10 Å. To pomeni, da je prostor nad aminokislino prazen in da je redko v interakciji. Po drugi strani pa ima levcin popolnoma drugačno porazdelitev. V povprečju je število aminokislin nad molekulo okrog 60, kar je povsem drugače kot pri nabitih aminokislinah, ki imajo največkrat 0 aminokislin v zgornji polkrogli. To pomeni, da se hidrofobne aminokisline zadržujejo v območjih, kjer nimajo proste površine.

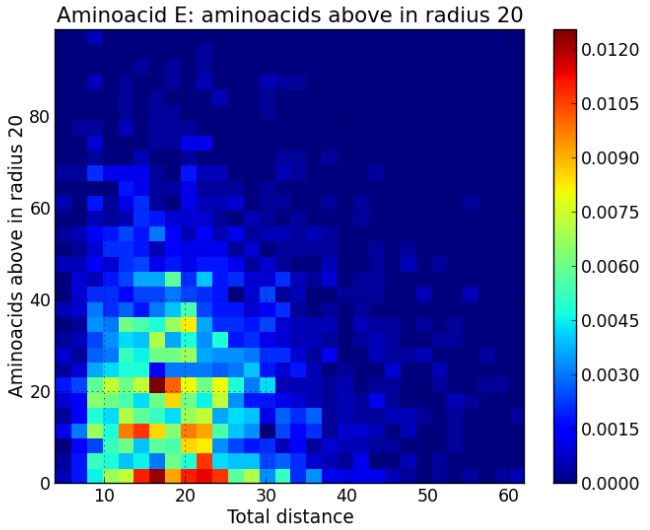
Z verjetnostnimi matrikami opišemo vse značilke in jih razlikujemo glede na aminokislino v paru. Pogostost pojavitve lahko direktno uporabimo pri računanju ocenjevalne funkcije.



Slika 5.7: Porazdelitev števila sosednjih aminokislin v polkrogli nad aminokislino lizin (K).



Slika 5.8: Porazdelitev števila sosednjih aminokislin v polkrogli nad aminokislino levcin (L).



Slika 5.9: Porazdelitev števila sosednjih aminokislin v polkrogli nad aminokislino glutamično kislino (E).

5.3 Ocenjevalne funkcije

V ocenjevalnih funkcijah uporabimo vrednosti značilk trenutne pozicije RNA in vrednosti značilk najbližjih aminokislin. Za dano značilko x in za dano trenutno razdaljo d po enačbi 5.1 dobimo frekvenco pojavitev iz verjetnostnih matrik. Verjetnostna matrika je specifična za vsako aminokislino posebej, zato je potrebno gledati v ustrezno izmed 20 matrik, ki predstavljajo trenutno opazovano aminokislino a.

$$f_x = P_a(x, d) \tag{5.1}$$

V naši oceni uporabimo več značilk, zato seštevamo več verjetnostnih matrik. Frekvence posameznih značilk seštejemo v uteženo vsoto X, ki predstavlja našo končno oceno pozicije. Vsako značilko utežimo, saj so nekatere pomembnejše in bolj vplivajo na uspešnost modela. Utežimo jih tako, da dobimo čim boljši napovedni model. Enačba 5.2 prikazuje računanje utežene vsote verjetnostnih matrik pozameznih značilk $(x_0, x_1, ..., x_N)$ in uteži teh

značilk $(w_0, w_i, ..., w_N)$.

$$X = \sum_{i=0}^{N} w_i P_a(x_i, d) \tag{5.2}$$

V vsaki iteraciji se izračuna značilke iz trenutne pozicije RNA v simulaciji in se jih primerja z verjetnostnimi matrikami, ki opisujejo obsoječe podatke 3D struktur kompleksov protein-RNA. Funkcija na sliki 5.10 prikazuje postopek izračuna ocene, glede na uteženo vsoto verjetnosti za različne značilke kandidata.

```
function evaluate(candidate, protein):
score ← 0
attributeList ← listOfAttributes()
weights ← initializeWeights(attributeList)
for attribute in attributeList do
   P ← probabilityMatrix(attribute)
   distance ← distanceToClosestAminoacid(candidate, protein)
   x ← P[candidate[attribute], distance]
   w ← weights[attribute]
   score ← score + x * w
end for
return score
```

Slika 5.10: Postopek ocenjevanja kandidata z uporabo verjetnostnih matrik.

Funkcija v vsaki poziciji izračuna lokalne 3D lastnosti glede na novo stanje. Iz tega stanja nato izračuna oceno, ki je odvisna od frekvence, določene značilke interakcije v realnih strukturah in razdalje med nukleotidom ter aminokislino. Postopek izračuna oceno pozicije. Boljšo oceno dosežemo, ko je verjetnost nove lokacije v učnih podatkih večja. Nato RNA premaknemo na najboljšo novo lokcijo. Podobno ocenimo tudi kvaliteto verige RNA tako, da s frekvenco pojavitve v učnih podatkih primerjamo trenutne razdalje med sosednjimi nukleotidi, kote med njimi in dolžine nukleotidov (razdalja med ribozo ter bazo nukleotida).

5.4 Uteževanje funkcij

Program dinamično določi uteži posameznih značilk tako, da večkrat požene simulacijo z različnimi utežmi značilk in na koncu simulacije izračuna evklidsko razdaljo med dejanskim položajem RNA v učnih podatkih in položajem RNA, ki ga dobimo v okviru umestitvenega algoritma. Uteži program nastavi tako, da čim bolj poveča natančnost napovedi. Za vsako značilko se med postopkom prilagajanja uteži izračunajo vrednosti, s katero se bo ocena značilke pomnožila. Postopek za prilagajanje uteži večkrat požene celoten umestitveni algoritem in izračuna razdaljo rešitev do dejanskih mest RNA. Vsako iteracijo se uteži spremenijo tako, da gre v smer boljših rešitev. Predpostavili smo, da so uteži neodvisne in v vsaki iteraciji spreminjali drugo utež. Algoritem prilagajanja uteži je prikazan s psevdokodo na sliki 5.11.

```
function fitWeights():
bestWeights ← null
score ← 0
weights ← initializeWeights()
while not stoppingCondition() do
  solution ← dockingAlgoritm(w)
  d = evaluateSolution(solution)
  if d > score then
    score ← d
    bestWeights ← w
  end if
  weights ← changeWeights(weights)
end while
return bestWeights
```

Slika 5.11: Postopek ocenjevanja kandidata z uporabo verjetnostnih matrik.

V ocenjevalno funkcijo smo vključili več različnih značilk, ki se po svoji vlogi razlikujejo na dva dela. Prvi del značilk je izračunan iz frekvence situacije, ki je odvisna od razdalje med aminokislino in nukleotidom. V algoritem

so vključene naslednje značilke:

Razdalja do proteina predstavlja frekvenco glede na razdaljo do proteina, podano v verjetnostni matriki. Nižje razdalje imajo ponavadi višjo frekvenco, zato ta funkcija pomaga približati strukturo k proteinu.

Razdalja do riboze podobno kot razdalja do proteina predstavlja frekvenco odvisno od razdalje do riboze. Skupaj z razdaljo do baze ta značilka pomaga pravilno obrniti nukleotid, ko pride blizu proteina, saj je pri interakcijah po navadi RNA obrnjen proti aminokislini s hrbtno stranjo verige.

Razdalja do baze predstavlja frekvenco razdalje do baze glede na relativno razdaljo med aminokislino in nukleotidom. Razdalja ima podobne vrednosti kot drugi dve razdalji. Razlika je pri aminokislinah, ki so v neposredni bližini nukleotida. Pri takih aminokislinah je baza verjetneje dlje od aminokisline kot riboza in se v optimizaciji obrne nukleotid stran od aminokisline.

Protein odboj je funkcija, ki prepreči, da bi se struktura preveč približala proteinu ali pa se premaknila v notranjost proteina. Odboj je prisoten samo na razdaljah, manjših od 4 Å.

Dolžino projekcije smo vključili zato, ker je ena izmed informativnejših značilk. Izračuana vrednost značilke in pogleda v verjetnostno matriko frekvenco pojavitve.

Kot aminokisline glede na projekcijo je druga izmed boljših značilk modela. Primer porazdelitve take značilke je prikazan na sliki 5.6.

V ocenjevalno funkcijo nismo vključili nekaterih obstoječih značilk, na primer števila aminokislin v zgornji polkrogli, saj ta značilka služi za razlikovanje aminokislin, ki so v centru proteina in tiste, ki so na obrobju proteina. Pri napovednem modelu je to potrebno, ker moramo oceniti vse aminokisline, tudi tiste, ki so v centru proteina. Ker naša metoda že vsebuje funkcijo za

odboj, te značilke nismo potrebovali, saj začne umestitveni algoritem na zunanji strani proteina in ne pride do notranjega dela.

V algoritem smo vključili še tri značilke, specifične za verigo RNA, ki niso odvisne od razdalje do proteina. Njihova naloga je ohranjanje strukture predlagane verige RNA skozi proces umestitve. Prevelika utež tem verigam močno omeji preiskovalni prostor in povzroča prehitro konvergenco, zato je bolje tolerirati netipične razdalje med sosednjimi nukleotidi, da dobimo ustreznejšo pozicijo. Premajhna utež povzroči, da se vsi nukleotidi približajo isti točki z boljšo skupno oceno.

Razdalja med sosednjimi nukleotidi predstavlja verjetnost, dane razdalje med sosednjima nukleotidoma. Ta značilka preprečuje, da bi vsi nukleotidi skonvergirali v isto točko, saj vrne boljšo oceno, če je razdalja med nukleotidi bližje povprečni. Na sliki 6.4 je prikazana porazdelitev dolžin nukleotida na vhodnih podatkih.

Kot med sosednjimi nukleotidi izračunamo po postopku, prikazanem v poglavju 4. Naloga te značilke je preprečevanje preostrih zavojev verige.

Dolžina nukleotida predstavlja verjetnost, da bo določen tip nukleotida (A,C,G,U) trenutno opazovane dolžine. Na sliki 6.5 je prikazana porazdelitev dolžin nukleotida na vhodnih podatkih. Ta značilka pomaga, da ostane baza oddaljena na realnejši razdalji od hrbtne strani RNA. Na sliki 6.5 je prikazana porazdelitev dolžin nukleotida na vhodnih podatkih.

5.5 Ocenjevanje kvalitete rešitev

Ocenjevalna funkcija koristi samo za vodenje iteracij preiskovalnega algoritma. Za vrednotenje algoritma, moramo rešitve primerjati z dejanskimi lokacijami interakcij v kompleksu protein-RNA. Najprej iz vsakega kompleksa izluščimo lokacije tistih RNA zaporedij, ki so v interakciji s proteinom. Definicija interakcije je predstavljena v poglavju 4. Poleg RNA zaporedij

vzamemo še dva, ki sta v neposredni bližini, da ohranimo sekvenco vsaj treh nukleotidov. Sledi računanje evklidske razdalje med vsakim nukleotidom predlagane rešitve in najbližjim nukleotidom, ki je še v interakciji s proteinom v realni strukturi. Ekvlidsko razdaljo med dvema točkama izračunamo po enačbi (5.3).

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \tag{5.3}$$

Tako dobimo za vsakega izmed petih nukleotidov razdalje, ki jih pretvorimo v končno oceno rešitve tako, da izračunamo njihovo povprečje.

Napisali smo tudi algoritem, ki naključno izbere koordinate iz notranjosti krogle, ki ima center v središču proteina in polmer polovico razdalje med najbolj oddaljenimi koordinatami. Algoritem generira 30 naključnih točk, na katerih postavi kratke verige. To je podobno kot pri generiranju začetnih točk, le da so tukaj točke lahko tudi v notranjosti proteina in se ne uporabljajo napovednega modela. Algoritem ocenimo z računanjem povprečne evklidske razdalje med rezultati umestitvenega algoritma do pravilne pozicije mesta RNA v interakciji.

Poglavje 6

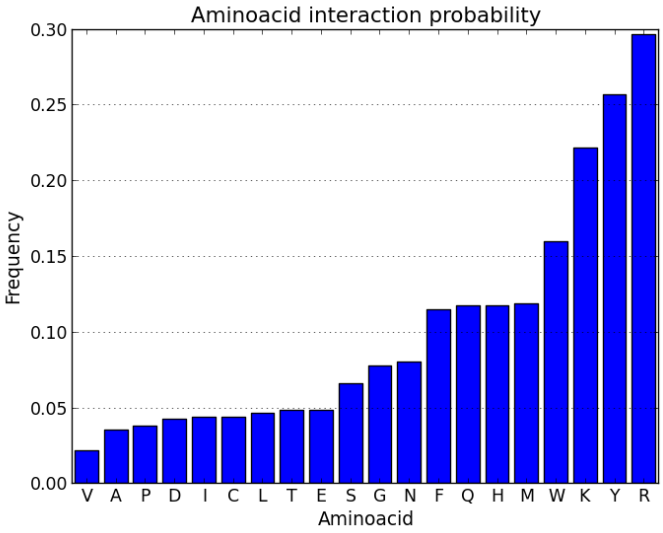
Rezultati

V okviru tega poglavja predstavimo podrobno računsko analizo 208 kompleksov protein-RNA. Predstavimo statistične podatke interakcij za aminokisline, nukleotide in njihove kombinacije. Nato predstavimo vpliv interakcij na različne tipe značilk na proteinu. Pokažemo rezultate napovednega modela za različne klasifikatorje in mere, med drugim površino AUC in koeficient MCC. Model primerjamo z večinskim klasifikatorjem in pokažemo, da naš model izboljša rešitve na dveh različnih množicah značilk. Na koncu še predstavimo rešitve umestitvenega algoritma, testiramo umestitveni algoritem z različnimi ocenjevalnimi funkcijami in primerjamo rešitve z realnimi podatki ter z modelom naključnega izbiranja.

6.1 Verjetnost interakcij aminokislin in nukleotidov

Pri analizi 3D strukture kompleksov protein-RNA smo ugotovili, da imajo nekatere aminokisline večjo težnjo, da vežejo RNA, med njimi arginin (R) in lizin (K). To se zgodi, ker je molekula RNA negativno nabita, omenjeni aminokislini pa imajo najbolj pozitiven naboj. Med pogostejšimi aminokislinami v interakciji so tudi nekatere daljše aminokisline, na primer triptofan (W) in tirozin (Y). Te aminokisline niso pozitivno nabite, vendar so lahko v interak-

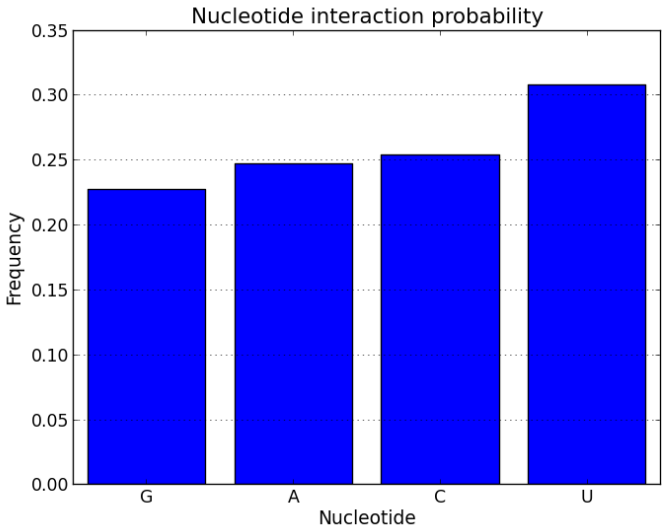
ciji, ker se lažje upognejo stran od RNA, medtem ko razdalja med atomom $C\alpha$ in nukleotidom ostane enaka. Slika 6.1 prikazuje frekvence aminokislin v interakciji.



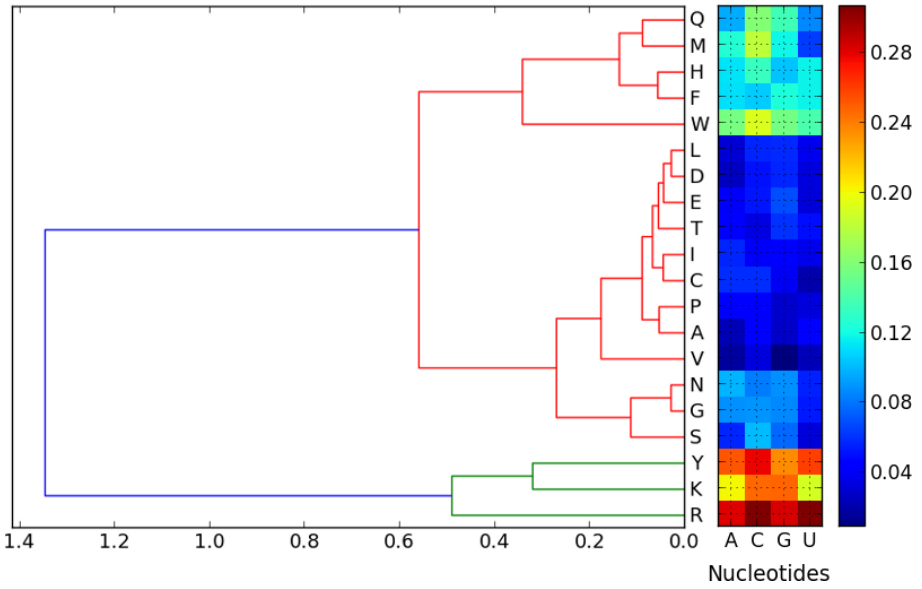
Slika 6.1: Pogostost aminokislin v interakciji (oddaljenost med aminokislino in najbližjim nukleotidom je manj kot 4 Å).

Slika 6.2 prikazuje frekvenco posameznih nukleotidov v interakciji. Nukleotidi so v interakciji s podobnimi verjetnostmi. Večje razlike vidimo šele, ko gledamo diagram kombinacij aminokislin in nukleotidov. Slika 6.3 prikazuje relativno pogostost interakcij za vsak par aminokisline in nukleotida. Uracil (U) ima rahlo višjo povprečno verjetnost interakcije. Vidimo tudi, da je uracil bolj prisoten pri argininu (R), ki je najpogostejše zastopan v interakcijah, in manj prisoten pri ostalih aminokislinah.

Aminokisline na sliki 6.3 so razvrščene glede na podobnost frekvence interakcij aminokislin. Podobnost interakcije med dvema aminokislinama izračunamo tako, da seštejemo kvadrate razlik med frekvencami pojavitev (vrstice) aminokislin pri različnih nukleotidih.



Slika 6.2: Pogostost nukleotidov v interakciji.

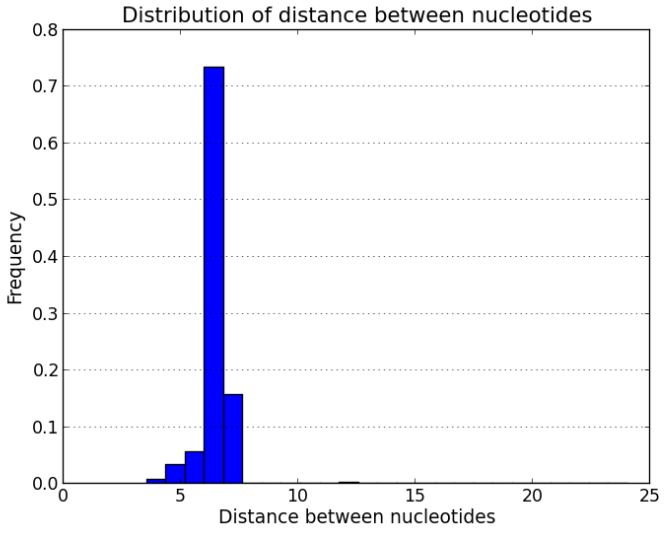


Slika 6.3: Pogostost interakcij nukleotidov je z vsako aminokislino utežena glede na vse pojavitve takih parov. Dendrogram prikazuje podobnost profilov aminokislin.

6.2 Porazdelitev značilk

6.2.1 Značilke na nukleotidu

Razdaljo med nukleotidi smo definirali kot razdaljo med sosednjima atomoma $C5$ na ribozi. Razdaljo je smiselno računati glede na atom $C5$, ker so nukleotidi preko tega atoma povezani med seboj in pride do najmanj variacij. Povprečna razdalja med nukleotidi je enaka 6.12 Å. Porazdelitev razdalj med nukleotidi je prikazana na histogramu 6.4.

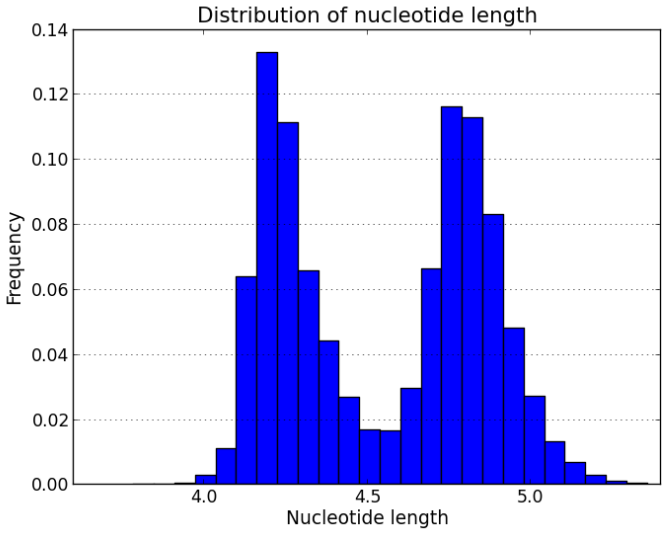


Slika 6.4: Porazdelitev razdalj med sosednjimi nukleotidi.

Dolžina nukleotida je razdalja med centrom baze in centrom riboze. S slike 6.5 je razvidno, da se dolžine pojavljajo v dveh območjih. To je zato, ker sta adenin in uracil krajši molekuli. Če pogledamo za vsak nukleotid posebej, je porazdelitev dolžine nukleotida podobna Gaussovi porazdelitvi. Povprečne dolžine nukleotidov so prikazane na tabeli 6.1.

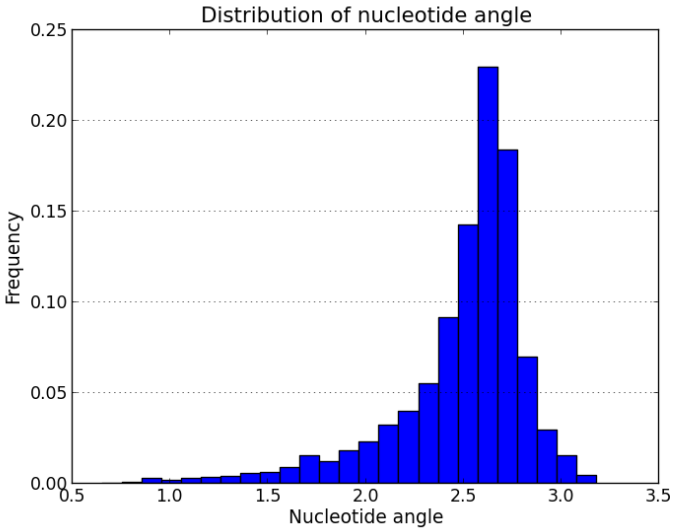
Tabela 6.1: Povprečne dolžine nukleotidov

Nukleotid	Dolžina
Adenin	4.752
Citozin	4.214
Gvanin	4.813
Uracil	4.254



Slika 6.5: Porazdelitev dolžin nukleotidov.

Kot med nukleotidi je opisan s tremi zaporednimi nukleotidi. Najpogosteje se pojavljajo topi koti. Ostri zavoji in ravne verige RNA so redke. Povprečen kot med nukleotidi je enak 2.45 radianov, kar ustreza 141 stopinjam.



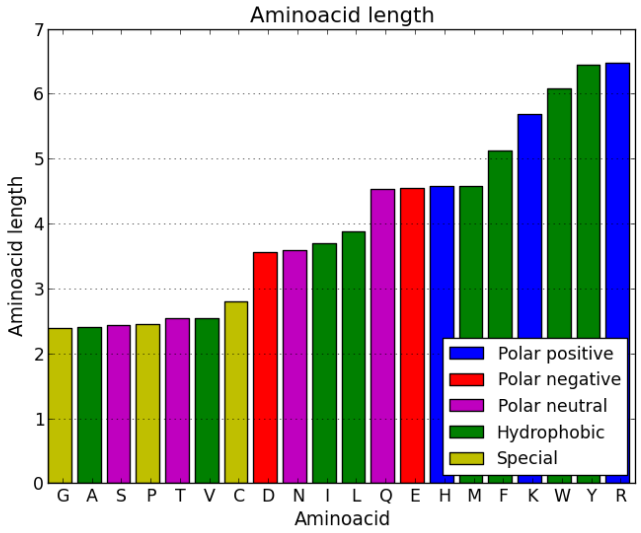
Slika 6.6: Porazdelitev kotov med sosednjimi nukleotidi.

6.2.2 Značilke na proteinu

Opisali smo razlike v povprečni vrednosti značilke pri vseh aminokislinah v primerjavi s povprečno vrednostjo značilke aminokislin, ki so v interakciji. Prikazani so histogrami, ki kažejo razliko, ki jo povzroči omejitev podatkov na interakcije. Za različne skupine aminokislin so prikazane tudi razlike, ki jih povzroči interakcija. Aminokisline razvrščamo v pet glavnih skupin:

1. polarne pozitivno nabite (A,H,R),
2. polarne negativno nabite (D,E),
3. polarne nevtralne (N,Q,S,T),
4. hidrofobne (A,F,I,L,M,V,W,Y),
5. posebni primeri (C,G,P).

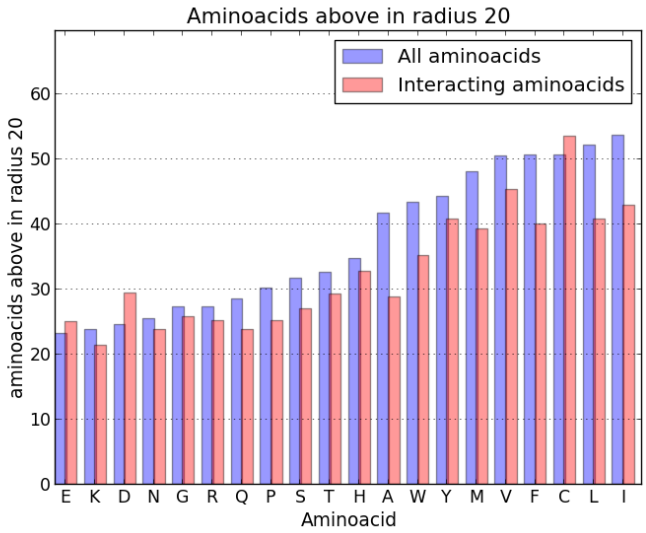
Dolžina aminokisline je ena izmed statičnih značilk aminokisline. Na sliki 6.7 so predstavljene dolžine aminokislin, obarvane glede na skupino, ki ji pripada.



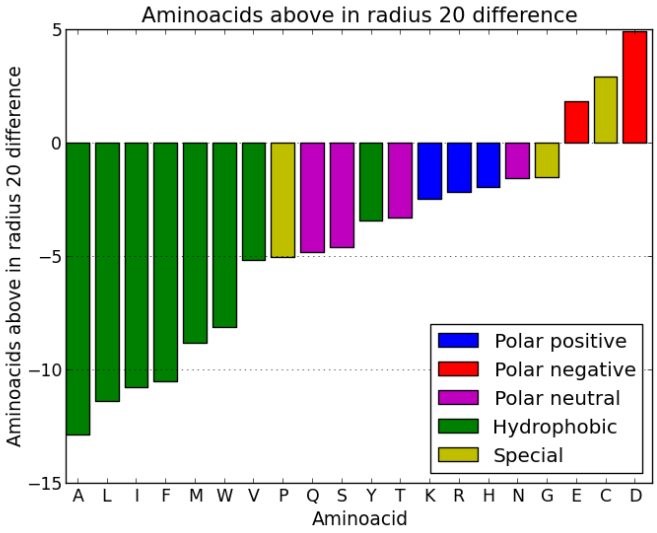
Slika 6.7: Dolžine aminokislin.

Število aminokislin v zgornji polkrogli smo testirali z različnimi velikostmi radija krogle. To število je pri hidrofobnih aminokislinah višje kot pri ostalih aminokislinah. To pomeni, da je več hidrofobnih aminokislin v notranjem delu proteina. Proteini se tako tipično oblikujejo zaradi prisotnosti vodnih molekul v okolici. Posledično je zato pri hidrofobnih aminokislinah, ki so v interakciji, večji padec števila okoliških aminokislin, saj aminokisline v notranjosti proteina ne morejo biti v interakciji. Negativno nabitim aminokislinam se pri interakcijah poveča število sosednjih aminokislin, ker se take aminokisline obrnejo proč od negativno nabite verige RNA, torej v notranjost proteina, kjer je večje število drugih aminokislin. Ta opažanja nakazujejo na to, da je ta značilka dober indikator lokacije aminokisline v molekuli. Povprečno število sosednjih aminokislin v zgornji polovici verige proteina prikazuje slika 6.8. Vpliv interakcije na število sosednjih aminokislin v zgornji

polkrogli je prikazan na sliki 6.9.



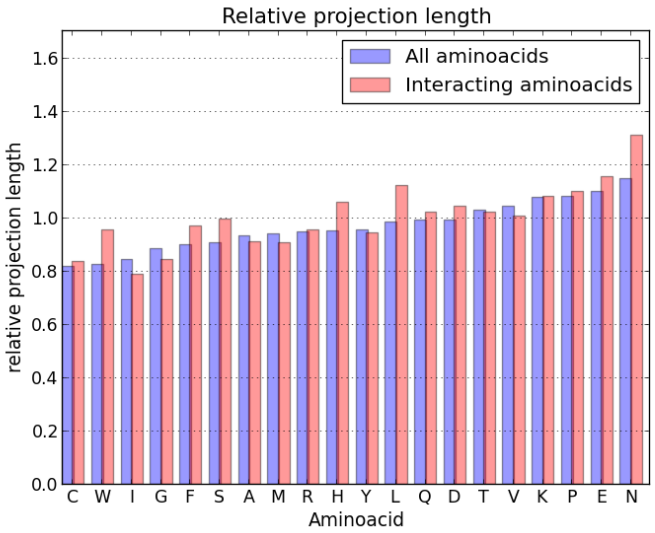
Slika 6.8: Povprečno število sosednjih aminokislin v polkrogli z radijem 20Å.



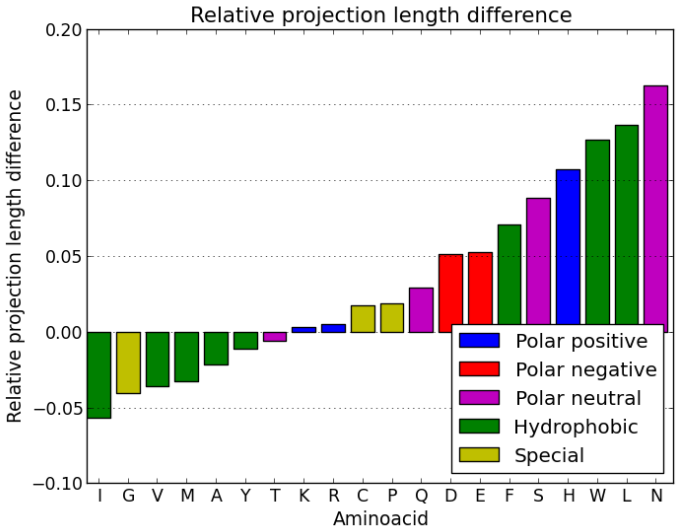
Slika 6.9: Vpliv interakcije na število sosednjih aminokislin v polkrogli z radijem 20Å.

Število aminokislin v krogli ima podobne lastnosti kot število aminokislin v zgornji polkrogli, saj sta značilki povezani. Vplivi so manjši, saj se število aminokislin v spodnji polkrogli manj spreminja pri interakcijah.

Dolžina projekcije aminokisline ponazarja, koliko se aminokislina nagiba stran od smernega vektorja n, ki ga definirajo atomi hrbtne verige C, N in $C\alpha$. Večje aminokisline se lahko bolj upogibajo ob interakciji, zato smo uvedli še značilko, kjer je dolžina projekcije normalizirana s povprečno dolžino aminokisline. Porazdelitev povprečnih verjetnosti relativne dolžine projekcije pri vseh aminokislinah in tistih, ki so v interakciji je ponazorjena na sliki 6.10. Razlika med vrednostmi relativne dolžine projekcije med vsemi aminokislinami in tistimi, ki so v interakciji je prikazana na sliki 6.11.



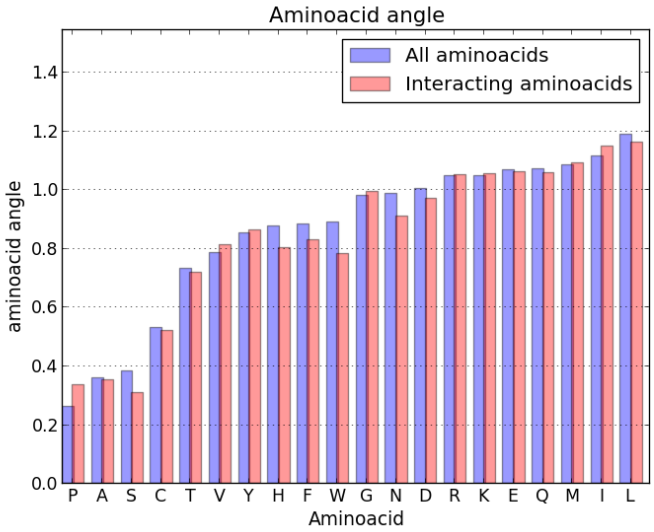
Slika 6.10: Relativna dolžina projekcije aminokisline.



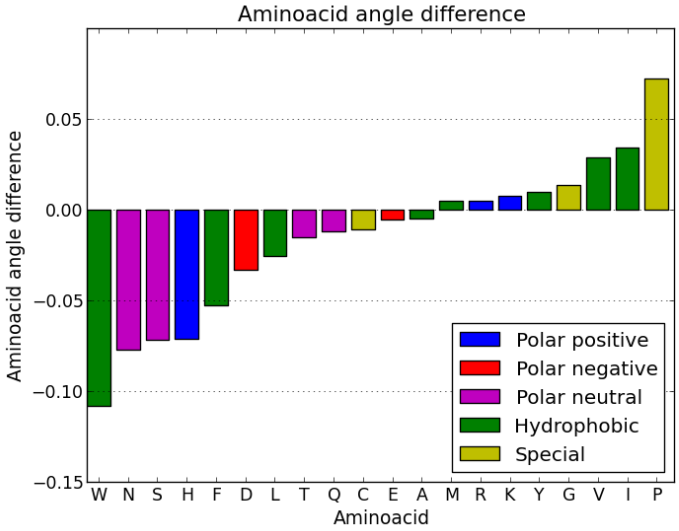
Slika 6.11: Razlika relativne dolžine projekcije aminokisline pri interakcijah.

Kot aminokisline glede na projekcijo se pri večjih aminokislinah lažje spreminja. Kot je prikazan na sliki 4.3 s simbolom α. Povprečna velikost kota pri vseh aminokislinah in tistih, ki so v interakciji je prikazana na sliki 6.12. Razlika med povprečnimi koti je prikazana na sliki 6.13.

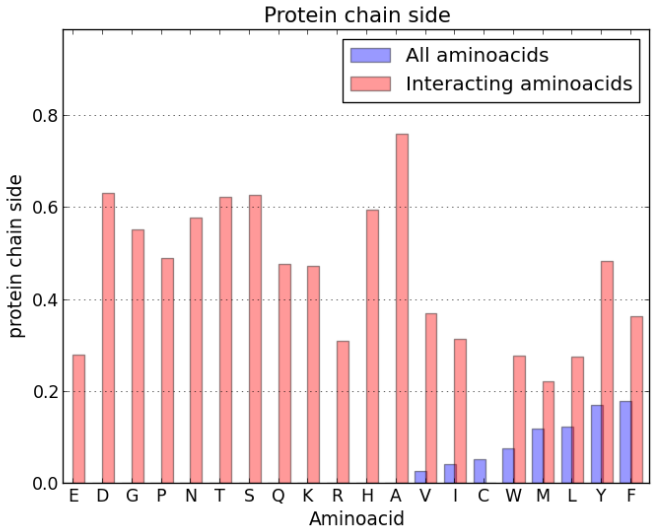
Stran aminokisline v interakciji predstavlja indikator, na kateri strani hrbtne verige aminokisline je nukleotid. Če je nukleotid na isti strani kot aminokislina, ima vrednost 1, na hrbtni strani pa ima vrednost -1. Ko te vrednosti povprečimo, dobimo število, ki pove število interakcij aminokisline, ki se nahajajo na isti strani verige kot aminokislina. Število 0 pomeni, da se obe situaciji pojavita z isto frekvenco. Slika 6.14 prikazuje vpliv interakcije na povprečno vrednost značilke, slika 6.15 pa prikazuje razliko med povprečnimi vrednostmi. Hidrofobne aminokisline imajo manjšo verjetnost, da bo interakcija na isti strani kot aminokislina. To se zgodi, ker se hidrofobne aminokisline obračajo navznoter, da pride do manj kontakta z molekulami vode.



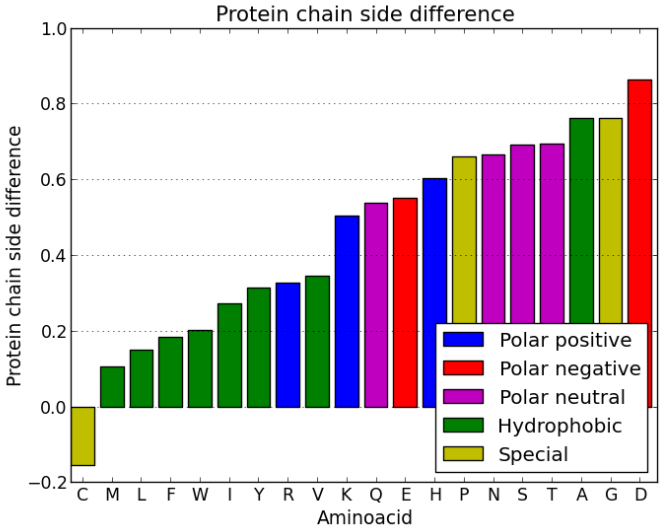
Slika 6.12: Povprečni kot vseh aminokislin in aminokislin v interakciji.



Slika 6.13: Razlika med povprečnimi koti aminokisline v interakciji.



Slika 6.14: Povprečna stran verige pri vseh aminokislinah ter aminokislinah v interakciji.



Slika 6.15: Vpliv interakcije na stran verige.

6.3 Rezultati napovednega modela

Zgradili smo model na podlagi referenčne verige proteina in referenčne verige RNA. Model na podlagi referenčne verige proteina sestavljajo podatki, ki za vsako aminokislino na verigi proteina vsebujejo podatke o najbližjem nukleotidu. Model na podlagi referenčne verige RNA je sestavljen iz podatkov, pri katerih za vsak nukleotid v verigi RNA, poiščemo najbližjo aminokislino. Oba modela uporabljata podatke o lokalni 3D strukturi proteina, podatke o lokalni 3D strukturi RNA in 3D značilke interakcij. Z značilkmi interakcij opisujemo kote med molekulama v interakciji in njihove relativne orientacije v prostoru. Uporabljene značilke so opisane v poglavju 4.

Za vrednotenje klasifikatorjev smo na 208 izbranih podatkovnih strukturah uporabili trikratno prečno preverjanje (dve tretjini učnih primerov in ena tretjina testnih primerov). Model smo vrednotili na klasifikacijskih drevesih, naivnem Bayesu, naključnih gozdovih ter SVM. Pri referenčni verigi proteina napovedni model doseže z uporabo klasifikacijskih dreves klasifikacijsko točnost 0.92, površino pod krivuljo ROC od 0.75 do 0.85 in MCC 0.42. Na podatkih referenčne verige RNA, imajo klasifikacijska drevesa klasifikacijsko točnost 0.81, površino pod krivuljo ROC od 0.77 do 0.82. Naivni Bayes ima površino pod krivuljo ROC od 0.75 do 0.78. Slabši rezultati izhajajo iz modelov zgrajenih samo na značilkah proteina, vendar so še vedno boljši od večinskega klasifikatorja. Iz tega sklepamo, da so lastnosti, ki se pojavijo med aminokislinam in nukleotidi v interakciji, značilne posebej za interakcije in se ne pojavljajo pri kombinaciji oddaljenih molekul. Rezultati so primerljivi z rezultati obstoječih metod, ki uporabljajo strukturne podatke za napovedovanje interakcij.

6.3.1 Model na podlagi 3D značilk strukture

Pri gradnji modela smo uporabili značilke, pridobljene iz 3D strukture kompleksa protein-RNA in jih uporabili pri gradnji modela. Na tabelah 6.2 in 6.3 so prikazane mere uspešnosti različnih klasifikatorjev.

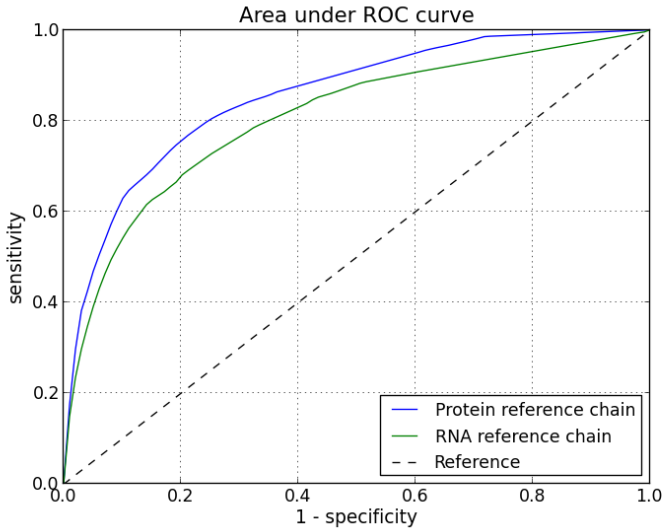
Tabela 6.2: Mere uspešnosti klasifikatorjev na podatkih referenčne verige proteina.

	Točnost	Preciznost	Priklic	AUC	MCC
Klasifikacijska drevesa	0.916	0.973	0.936	0.854	0.420
Naivni Bayes	0.862	0.900	0.946	0.825	0.343
Naključni gozdovi	0.906	1.000	0.906	0.840	0.002
SVM	0.906	0.999	0.908	0.865	0.090
Večinski klasifikator	0.906	0.000	0.000	0.500	0.000

Tabela 6.3: Mere uspešnosti klasifikatorjev na podatkih referenčne verige RNA.

	Točnost	Preciznost	Priklic	AUC	MCC
Klasifikacijska drevesa	0.811	0.922	0.839	0.806	0.461
Naivni Bayes	0.765	0.841	0.843	0.784	0.385
Naključni gozdovi	0.776	0.989	0.773	0.816	0.297
SVM	0.786	0.932	0.804	0.802	0.419
Večinski klasifikator	0.744	0.000	0.000	0.500	0.000

Slika 6.16 prikazuje krivuljo ROC modela klasifikacijskih dreves, zgrajenega na vseh značilkah, razen tistih, ki definirajo razdalje med aminokislinami in nukleotidi. Površina pod krivuljo ROC je večja pri modelu referenčne verige proteina. To se zgodi, ker je več primerov, kjer ne pride do interakcij kot na referenčni verigi RNA. Tako je večinski klasifikator uspešnejši na proteinskem modelu, ker napove vse primere negativno. Zaradi podobnih razlogov lahko tudi pri drugih klasifikatorjih dobimo večje točnosti.



Slika 6.16: Krivulji ROC na podlagi referenčne verige proteina in na podlagi referenčne verige RNA.

6.3.2 Model na podlagi 3D značilk proteina

Včasih imamo na voljo le 3D podatke o proteinu, za RNA pa imamo na voljo le zaporedje nukleotidov. Zato je smiselno zgraditi še model, ki se ne zanaša na značilke, pridobljene iz 3D struktur nukleotida. Posledično ne moremo uporabiti tudi značilk, pridobljenih iz interakcij, kot so na primer stran verige proteina, kjer se nahaja nukleotid. Zgrajen model smo testirali na podoben način kot model, ki vsebuje tudi ostale značilke. Napovedni model na osnovi strukturnih podatkov proteina ima površino pod krivuljo ROC do 0.1 slabšo, kljub temu ima model višjo stopnjo klasifikacijske točnosti od večinskega klasifikatorja razen pri naivnemu Bayesu.

Tabela 6.4 predstavlja uspešnost različnih napovednih modelov in večinskega klasifikatorja na podlagi referenčne verige proteina.

Tabela 6.5 predstavlja uspešnost klasifikatorja, ki uporabi strukturne značilke proteina za napovedovanje interakcij, zgrajenih na podlagi referenčne verige RNA. Na tej verigi vsak nukleotid porabimo samo enkrat, aminokisline pa se lahko v modelu uporabijo večkrat.

Tabela 6.4: Mere uspešnosti klasifikatorjev, osnovanih na značilkah o proteinu, glede na referenčno verigo proteina.

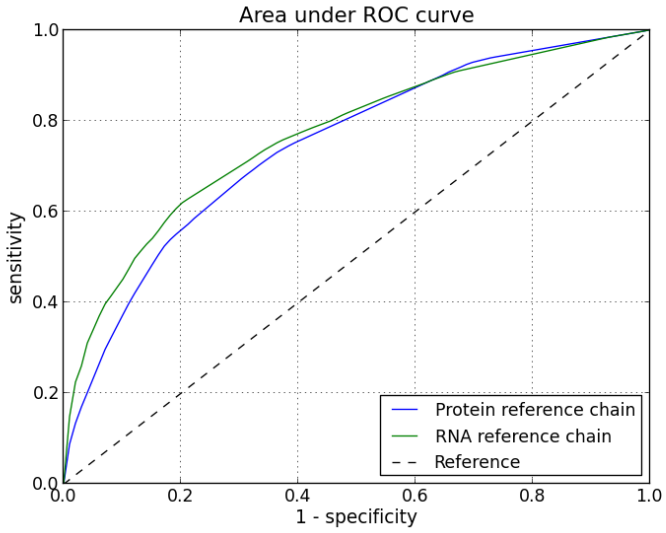
	Točnost	Preciznost	Priklic	AUC	MCC
Klasifikacijska drevesa	0.907	0.997	0.909	0.756	0.143
Naivni Bayes	0.845	0.900	0.946	0.716	0.222
Naključni gozdovi	0.906	0.999	0.906	0.750	0.037
SVM	0.903	1.000	0.903	0.678	0.000
Večinski klasifikator	0.906	0.000	0.000	0.500	0.000

Tabela 6.5: Mere uspešnosti klasifikatorjev, osnovanih na značilkah o proteinu, glede na referenčno verigo RNA.

	Točnost	Preciznost	Priklic	AUC	MCC
Klasifikacijska drevesa	0.791	0.948	0.806	0.767	0.378
Naivni Bayes	0.762	0.864	0.823	0.752	0.356
Naključni gozdovi	0.788	0.797	0.977	0.782	0.359
SVM	0.761	0.943	0.774	0.767	0.325
Večinski klasifikator	0.744	0.000	0.000	0.500	0.000

Slika 6.17 prikazuje krivulji ROC modela klasifikacijsih dreves, ki uporablja značilke proteina. Z modro barvo je prikazana krivulja ROC za referenčno verigo proteina, z zeleno pa krivulja ROC za referenčno verigo RNA. Napovedna točnost proteinske verige je tukaj slabša od napovedne točnosti verige RNA. Pri referenčni verigi RNA so v učni množici izbrane aminokisline v interakciji, pri referenčni verigi proteina pa vse aminokisline in samo nukleotidi v interakciji. Na podlagi samih podatkov o proteinu je težko napovedati, kateri nukleotid bo v interakciji, saj o njih nimamo podatkov. Pri referenčni verigi RNA je nukleotid že določen, priredimo mu le aminokislino, za katero lažje napovemo ali bo prišlo do interakcije, saj lahko to sklepamo

iz učnih podatkov o proteinu.



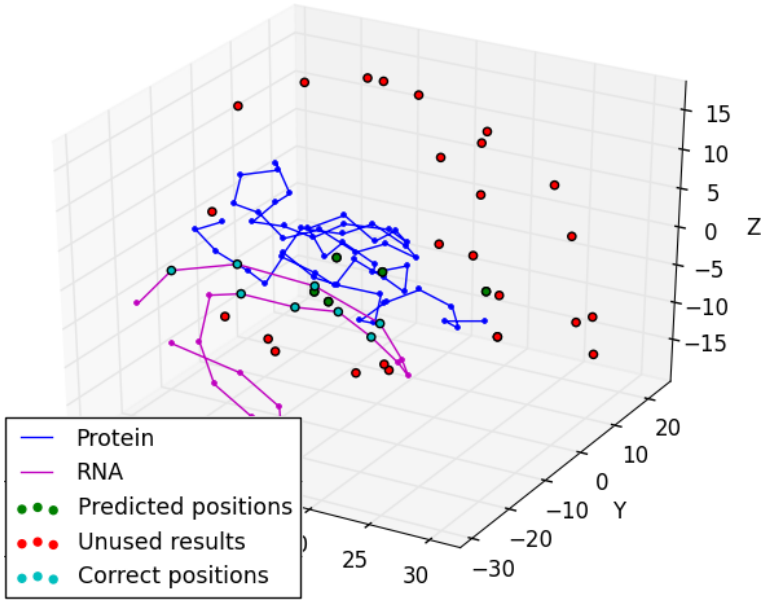
Slika 6.17: Krivulji ROC na podlagi referenčne verige proteina in na podlagi referenčne verige RNA, osnovane na značilkah o proteinu.

6.4 Rezultati umestitvenega algoritma

Umestitveni algoritem smo ovrednotili tako, da smo izračunali evklidsko razdaljo med položajem RNA, ki je rezultat umestitvenega algoritma in dejansko lokacijo RNA iz 3D podatkov kompleksa, ki vstopa v interakcijo skupaj z njegovimi sosedi. Z uporabo kombinacije najuspešnejših uteži, prikazanih v tabeli 6.7, smo testirali naš algoritem na 208 izbranih strukturah. Vzeli smo pet najboljših rešitev iz vsakega primera, nato smo izračunali povprečno razdaljo do dejanskega mesta interakcije.

Slabost našega umestitvenega algoritma je ta, da algoritem v veliko primerih prehitro skonvergira v lokalni optimum. Zato smo za vsako strukturo izmed 30 iteracij izračunali sredinsko točko naše sekvence in izbrali samo tiste, ki so dovolj blizu proteina. Pri ostalih pa zaključimo, da je optimizacija neuspešna. Slika 6.18 prikazuje lokacije predlaganih rešitev umestitvenega

algoritma za strukturo proteina *1a1t*. Z modro je označena osnovna veriga proteina, z vijolično barvo pa osnovna veriga RNA.



Slika 6.18: Lokacije rešitev umestitvenega algoritma so označene z zeleno, lokacije neuspešnih iteracij so označene z rdečo, veriga proteina je označena z modro, veriga RNA pa z vijolično barvo.

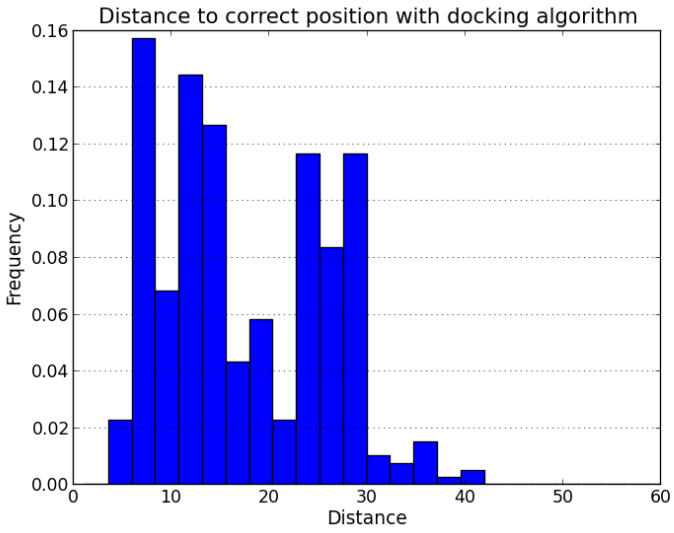
Na sliki 6.6 zelene točke predstavljajo filtrirane lokacije glede na bližino proteina. Z rdečo barvo so označene vse rešitve, ki smo jih odstranili zaradi prevelike oddaljenosti. Pravilna mesta interakcije so označene s svetlo modrimi točkami. Kot vidimo, ima algoritem velik upadek rešitev, kar je posledica ujetja v lokalnih ekstremih preiskovalnega prostora. S filtriranjem izkoristimo lastnost, da v primerih, ko optimizacija uspe pripeljati strukturo do proteina, je ta natančnejša od metode naključnega vzorčenja. Vrednosti povprečne razdalje med rešitvami algoritma in rešitvami primerov naključnega vzorčenja so prikazane v tabeli 6.6.

Povprečna razdalja med rezultatom algoritma in dejansko lokacijo interakcij znaša 16.23 Å, kar je več kot metoda naključnega vzorčenja, ki doseže povprečno bližino realnega stanja 19.79 Å. Porazdelitev kvalitete rešitev je prikazana na sliki 6.19. Naključno vzorčenje izgubi natančnost, med drugim

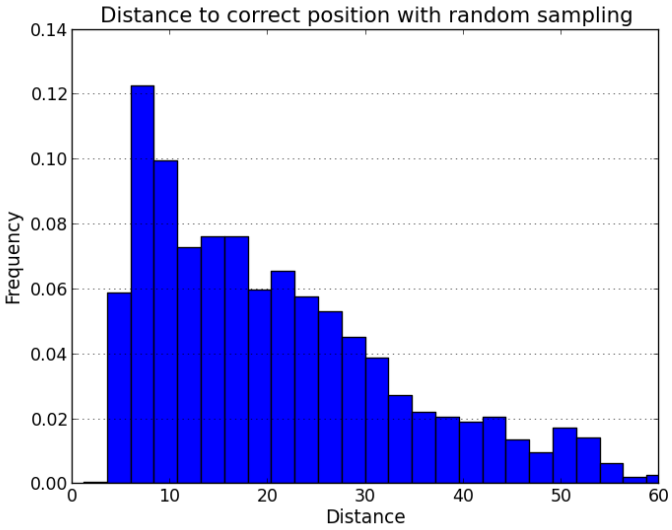
Tabela 6.6: Primerjava uspešnosti algoritmov. μ predstavlja povprečno razdaljo do optimalne rešitve, σ pa standardni odklon.

Algoritem	μ	σ
Umestitveni algoritem	16.23	8.42
Naključno vzorčenje	19.79	13.34

zaradi tega, ker velikokrat izbere rešitve znotraj proteina, ki se ne pojavijo na realnih strukturah. Slika 6.20 prikazuje porazdelitev razdalj algoritma naključnega vzorčenja do pravilnega mesta interakcije.



Slika 6.19: Porazdelitev razdalj rešitev algoritma do pravilne lokacije mest interakcije.

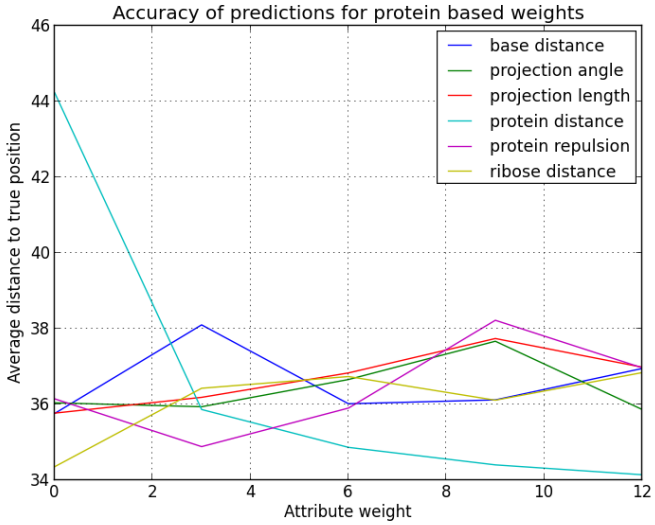


Slika 6.20: Porazdelitev razdalj rešitev naključnega vzorčenja do pravilne lokacije mest interakcije.

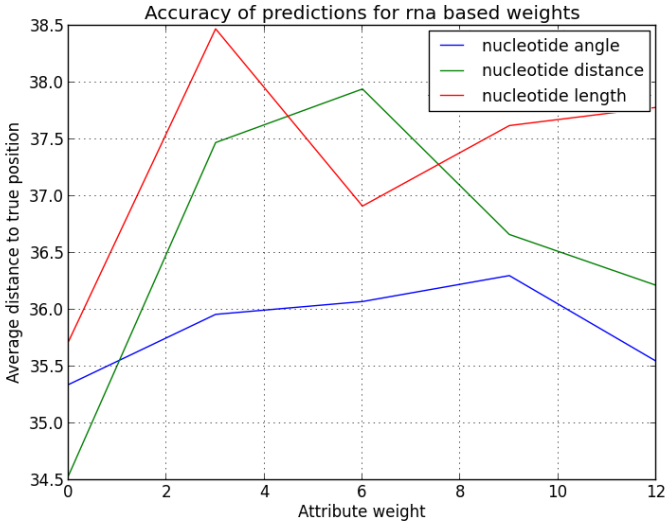
6.4.1 Spreminjanje uteži

Uteži funkcij določajo, v kolikšni meri bo vsaka značilka vključena v oceno. Za ugotavljanje bolj informativnih značilk smo testirali umestitveni algoritem z različnimi utežmi. Na sliki 6.21 je prikazana povprečna uspešnost več poskusov optimizacije za različne vrednosti značilk proteina in razdalj med proteinom in RNA. Na sliki 6.22 je prikazana povprečna uspešnost več poskusov optimizacije za vrednosti uteži, ki določajo pomembnost ohranjanja razdalje med nukleotidi, koti med nukleotidi in dolžine nukleotidov.

V tabeli 6.7 je seznam uteži, ki so se izkazale za najboljše. Te uteži pridobimo iz dinamičnega prilagajanja, njihovo uspešnost lahko vidimo na slikah 6.21 in 6.22. Uteži v tabeli 6.7 smo uporabili za iskanje rešitev, ki smo jih filtrirali glede na bližino proteina in nato ovrednotili glede na pravilno pozicijo mesta interakcije ter primerjali povprečno dolžino z metodo naključnega vzorčenja.



Slika 6.21: Uspešnost optimizacije pri različnih utežeh značilk proteina.



Slika 6.22: Uspešnost optimizacije pri različnih utežeh značilk RNA.

Tabela 6.7: Značilke in uteži, ki so bile uporabljene za ocenjevanje umestitvenega algoritma.

Tip značilke	Značilka	Utež
Interakcija	razdalja do riboze	12.0
Interakcija	razdalja do baze	3.0
Interakcija	razdalja do proteina	1.0
Protein	dolžina projekcije	9.0
Protein	kot projekcije	9.0
Protein	odboj proteina	9.0
Rna	dolžina nukleotida	3.0
Rna	kot med sosednjimi nukleotidi	9.0
Rna	razdalja med sosednjimi nukleotidi	6.0

Poglavje 7

Sklepi

Analiza velike količine strukturnih podatkov zahteva učinkovite računske metode napovedovanja interakcij protein-RNA. V okviru magistrskega dela smo razvili računsko metodo za analizo podatkov, ki vključuje strojno učenje.

Predstavili smo problem napovedovanja interakcij protein-RNA. Opisali smo obstoječe metode, ki uporabljajo zaporedje proteina in RNA ter metode, ki uporabljajo strukturne lastnosti. Predstavili smo tudi obstoječe metode na problemu umestitve RNA in proteina.

Določili smo verjetnost interakcij posameznih nukleotidov in aminokislin. Verjetnosti interakcije nukleotidov s proteinom so zelo podobne. Aminokisline pa kažejo veliko raznolikost v interakciji z RNA. V interakcijo z RNA vstopajo posamezne aminokisline in ne njihova osnovna veriga. V interakcijo z aminokislino vstopa negativno nabita osnovna veriga RNA, bistveno manj pa same dušikove baze.

Opazovali smo lokalne geometrijske spremembe pri parih, ki so v interakciji. Opazili smo spreminjanje strukture nekaterih aminokislin, ko se približa RNA, in višjo prisotnost hidrofobnih aminokislin v notranjosti proteina.

Prispevki magistrskega dela so metode, ki za dane strukture proteina in RNA določijo mesta, pri katerih pride do interakcije. Pomemben prispevek je definicija in uporaba novih strukturnih značilk.

Zgrajeni napovedni model dosega visoko natančnost napovedi mesta inte-

rakcij (tabeli 6.2 in 6.4 prikazujeta natančnosti našega modela). Za napoved s klasifikacijskimi drevesi znaša površina pod krivuljo ROC 0.85 in koeficient MCC 0.42, kar je primerljivo z natančnostjo najboljših obstoječih metod. Tabela 2.1 prikazuje uspešnosti obstoječih metod.

Drugi prispevek tega dela je ocenjevalna funkcija, ki je prilagojena problemu in uporablja napovedni model za generiranje začetnih pozicij. Uporablja verjetnosti iz empiričnih podatkov za vodeno kombinatorično preiskovanje najboljše umestitve kratke verige RNA s proteinom.

Predlagani pristop ima veliko možnosti izboljšav. Lahko bi ga izboljšali z dodajanjem novih značilk, kar zahteva dobro poznavanje biokemijskih in fizikalnih lastnosti proteina in RNA. Izpeljava možnih dodatnih značilk bi vključevala uporabo primerjave 3D strukture proteina in RNA v interakciji.

Strukturne lastnosti bi dodatno opisali z navajanjem sekundarnih struktur, kot na primer vijačnica alfa in ploskev beta. Vključili bi lahko tudi značilke za opisovanje krajših zaporedij aminokislin ali nukleotidov.

Umestitveni algoritem lahko ločimo na dve fazi. Prva faza je računanje samo z eno točko, ki ga potem spremenimo v računanje lokalne strukture zaporedja nukleotidov v drugi fazi algoritma. Spreminjanje RNA med iteracijami omeji prostor in povzroči hitrejšo konvergenco. Poleg tega bi v ocenjevalno funkcijo lahko dodali še druge značilke, upoštevali vodikove vezi ali vezi med posameznimi atomi in na ta način izboljšali trenutni pristop, ki deli nukleotid le na ribozo in dušikovo bazo.

Literatura

[1] R. Abagyan, M. Totrov, D. Kuznetsov, "ICM-a new method for protein modeling and docking." J Comput Chem 20, str. 412–427, 1999.

[2] C. A. Baxter et. al, "Flexible docking using Tabu search and an empirical estimate of binding affinity", Proteins: Structure, Function, and Genetics 33, str. 367–382, 1998.

[3] H. Bohm, "LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads", J Comput Aided Mol Des 6, str. 593–606, 1992.

[4] F. Campeotto, A. D. Palu, A. Dovier, F. Fioretto, E. Pontelli, "A Contstraint Solver for Flexible Protein Models", Journal of Artificial Intelligence Research 48, str. 953–1000, 2013.

[5] C. W. Cheng, E. C. Su, J. K. Hwang, T. Y. Sung, L. W. HSU, "Predicting RNA-binding sites of proteins using support vector machines and evolutionary information", BMC Bioinformatics 9 (suppl. 12), S6, 2008.

[6] D. Cirillo, F. Agostini, in G. G. Tartaglia, "Predictions of protein–RNA interactions," WIREs Comput Mol Sci 3, str. 161–175, 2013.

[7] D. J. Diller, K. M. Merz, "High throughput docking for library design and library prioritization", Proteins: Structure, Function, and Genetics 43, str. 113–124, 2001.

[8] C. Dominiquez, R. Boelens, A. M. Bonvin, "HADDOCK: a protein-protein docking approach based on biochemical or biophysical information", J Am Chem Soc 125, str. 1731–1737, 2003.

[9] T. J. A. Ewing, S. Makino, A.G. Skillman, I.D. Kuntz, "DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases", J Comput Aided Mol Des 15, str. 411–428, 2001.

[10] M. Fernandez, Y. Kumagai, D. M. Standley, A. Sarai, K. Mizuguchi, S. Ahmad, "Prediction of dinucleotide-specific RNA-binding sites in proteins", MBC Bioinformatics 12 (suppl 13), S6, 2011.

[11] K. Fujishima, M. Komasa, S. Kitamura, H. Suzuki, M. Tomita, A. Kanai, "Proteom-wide prediction of novel DNA/RNA-binding proteins using amino acid-composition and periodicity in the hyperthermophilic archaeon Pyrococcus furiosus", DNA Res 14, str 91–102, 2007.

[12] T. Fukunaga, H. Ozaki, G. Terai, K. Asai, W. Iwasaki, H. Kiryu, "CapR: Revealing structural specificites of RNA-binding protein target recognition using CLIP-seq data", Genome Biology 15, 2014.

[13] H. A. Gabb, R. M. Jackson, M. J. Sternberg, "Modelling protein docking using shape complementarity, electrostatics and biochemical information.", J mol Biol 272, str. 106–120, 1997.

[14] T. Hart, R. Read, "A multiple-start Monte Carlo docking method", Proteins: Structure, Function and Genetics 13, str. 206–222, 1992.

[15] G. Jones et. al., "Development and validation of a genetic algorithm for flexible docking", J Mol Biol 267, str. 727–748, 1997.

[16] S. Jones, D. T. A. Daley, N. M. Luscombe, H. M. Berman, J. M. Thornton, "Protein-RNA interactions: a structural analysis", Nucleic Acids Research 29:4, str. 943–954, 2001.

[17] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, I. A. Vakser, "Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques", Proc Natl Acad Sci USA 89, str. 2195–2199, 1992.

[18] O. T. Kim, K. Yura, N. Go, "Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction," Nucleic Acids Res 34, str. 6540–6560, 2006.

[19] B. Kramer, M. Rarey, T. Lengauer, "Evaluation of the FlexX incremental construction algorithm for protein-ligand docking", Proteins: Structur, Function and Genetics 37, str. 228–241, 1999.

[20] M. Kumar, M. M. Gromiha, G. P. S. Raghava, "SVM based prediction of RNA-binding proteins using binding residues and evolutionary information", J Mol Recognit 22, str. 303–313, 2011.

[21] X. Li, H. Kazan, H. D. Lipshitz, Q. D. Morris, "Finding the target sites of RNA-binding proteins", WIREs RNA 5, str. 111–130, 2014.

[22] M. Liu, S. Wang, "MCDOCK: a Monte Carlo simulation approach to the molecular docking problem", J Comput Aided Mol Des 13, str. 435–451, 1999.

[23] T. Liu, X. Geng, X. Zheng, R. Li, J. Wang, "Accurate prediction of protein structural class using auto covariance transformation of PSI-LAST profiles", Amino Acids 42, str 2243–2249, 2012.

[24] W. Luo, J. Pei, Y. Zhu, "A fast protein-ligand docking algorithm based on hydrogen bond matching and surface shape complementarity", J Mol Model 16, str. 903–913, 2010.

[25] X.Ma, J. Guo, J. Wu, H. Liu, J. Yu, J. Xie, X. Sun, "Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature", Proteins 79, str. 1230–1239, 2011.

[26] S. R. Maetschke, Z. Yuan, "Exploiting structural and topological information to improve prediction of RNA-protein binding sites", BMC Bioinformatics 10, str. 341, 2009.

[27] C. McMartin, R. Bohacek, "QXP: powerful, rapid computer algorithms for structure-based drug design", J Comput Aided Mol Des 11, str. 333–344, 1997.

[28] M. Miller, S. Kearsley, D. Underwood, R. Sheridan, "FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure", J Comput Aided Mol Des 8, str. 153–174, 1994.

[29] S. Miller, J. Janin, A. M. Lesk, C. Chothia, "Interior and surface of monomeric proteins", J Mol Biol 196, 641–656, 1987.

[30] M. Mizutani, N. Tomioka, A. Itai, "Rational automatic search method for stable docking models of protein and ligand", J Mol Biol 243, str. 310–326, 1996.

[31] G.M. Morris et. al., "Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function", J Comput Chem 19, str. 1639–1662, 1998.

[32] U. Muppirala, V. Honovar, D. Dobbs, "Predicting RNA-protein interactions using only sequence information", BMC Bioinformatics 12, str. 489, 2011.

[33] L. Nilsson, M. Karplus, "Empirical energy functions for energy minimization and dynamics of nucleic acids Supported in part by a giant from the national institutes of health", J Comput Chem 7, str. 591–616, 1986.

[34] L. Perez-cano, A. Solernou, C. Pons, J. Fernandez-Recio, "Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potetntials", Pac Symp Biocomput 293, str. 301, 2007.

[35] L. Perez-Cano, J. Fernandez-Recio, "Optimal protein-RNA area, OPRA: a propensity-based methods to identify RNA-binding sites on proteins," Proteins 78, str. 25–35, 2010.

[36] T. Puton, L. Koslowski, I. Tuszynska, K. Rother, J. M. Bujnicki, "Computational methods for prediction of protein-RNA interactions", J Struct Biol 179, str. 261–68, 2012.

[37] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, "A fast flexible docking method using an incremental construction algorithm", J Mol Biol 261, str. 470–489, 1996.

[38] D. W. Ritchie, G. J. Kemp, "Protein docking using spherical polar Fourier correlations", Proteins 39, str. 178–194, 2000.

[39] A. Sacan, O. Ozturk, H. Ferhatosmanoglu, Y. Wang, "LFM-Pro: a tool for detecting significant local structural sites in proteins", Bioinformatics 23, str. 709–716, 2007.

[40] D. Schneiderman-Duhovny, Y. Inbar, R. Nussinov, H. J. Wolfson, "PatchDock and SymmDock: servers for rigid and symmetric docking", Nucleic Acids Res 33, str. 363–367, 2005.

[41] S. Shazman, Y. Mandel-Gutfreund, " Classifying RNA-binding proteins based on electrostatic properties", PLoS Comput Biol 4, 2008.

[42] R. M. Sweet, D. Eisenberg, "Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure", J Mol Biol, 171, str. 479–488, 1983.

[43] P. Tao, L. Lai, "Protein ligand docking based on empirical method for binding affinity estimation", J Comput Aided Mol Des 15, str. 429–446, 2001.

[44] J. Taylor, R. Burnett, "DARWIN: a program for docking flexible molecules", Proteins: Structure, Function and Genetics 41, str. 173–191, 2000.

[45] M. Terribilini, J .D. Sandler, J-H. Lee, P. Zaback, R.I Jernigan, V. Honovar, D. Dobbs, "RNABindR: a server of analyzing and predicting RNA-binding sites in proteins", Nucleic Acids Res 35, str. 578–584, 2007.

[46] F. Towfic, C. Caragea, D. C. Gemperline, D. Dobbs, V. Honavar, "Struct-NB: predicting protein-RNA binding sites using structural features", Int J Data Min Bioinform 4, str. 21–43, 2010.

[47] J. Trosset, H. Scheraga, "Prodock: software package for protein modeling and docking", J Comput Chem 20, str. 412–427, 1999.

[48] G. Trinquier, Y.H. Sanejouand, Protein Eng.11, str. 153–169, 1998.

[49] I. Tuszynska, J. M. Bujnicki, "DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking", BMC Bioinformatics 12, str. 348, 2011.

[50] M. Verdonk et. al, "Improved protein-ligand docking using GOLD", Proteins 52, str. 609–623, 2003.

[51] C. M. Ventkatachalam, X. Jiang, T. Oldfield, M. Waldman, "LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites", J Mol Graph Model 21, str. 289–307, 2003.

[52] R. X. Wang, L. H. Lai, S. M. Wang, "Further development and validation of empirical scoring functions for structure-based binding affinity prediction", J Comput Aided Mol Des 16, str. 11–26, 2002.

[53] Y. Wang, Z. Xue, G. Shen, J. Xu, "PRINTR: prediction of RNA binding sites in proteins using SVM and profiles", Amino Acids 35, str. 295–302, 2008.

[54] L. Wang, S. J. Brown, "BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences", Nucleic Acids Res 34, str. 243–248, 2006.

[55] P. Weiner, P. Kollman, "AMBER: assisted model building with energy refinement. A general program for modeling molecules and their interactions", J Comput Chem 2, str. 287–303, 1981.

[56] W. Welch, J. Ruppert, A. Jain, "Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites", Chem Biol 3, str. 449–462, 1996.

[57] X. Yu, J. Cao, Y. Cai, T. Shi, Y. Li, "Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines", J Theor Biol 240, str. 175–184, 2005.

[58] H. Zhao, Y. Yang, Y. Zhou, "Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets", Nucleic Acids Res 39, str. 3017–3025, 2011.

[59] S. Zheng, T. A. Robertson, G. Varani, "A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins", FEBS J 274, str. 6378–6391, 2007.

[60] H. Zhao, Y. Yang, Y. Zhou, "Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction", RNA Biol 8, str. 988–996, 2011.

[61] Protein Data Bank (2014). Dostopno na: `http://www.rcsb.org/`